Four Settings and a Proposal (For the Exploration of Value Alignment in AI)

Pablo Noriega[1*†] and Enric Plaza[1*†]

[1]Artificial Intelligence Institute, Spanish Scientific Research Council (IIIA-CSIC), CAMPUS UB, Barcelona, 08193, Catalonia, Spain.

*Corresponding author(s). E-mail(s): pablo@iiia.csic.es; enric@iiia.csic.es;
†These authors contributed equally to this work.

Abstract

We outline four problem settings that illustrate the type of distinctive conceptual, methodological and empirical challenges involved in the design of AI systems that are aligned with human values. The four settings complement each other but each one is meant to reveal some specific aspects of the overall problem. The settings are: alignment on collective coordination, alignment of individual actions, an experimental approach to the engineering of value alignment and the empirical problems of making value alignment work in practice.
This discussion is meant to provide substance for the exploration of the interplay of the notions of autonomy, governance and values that is motivated by the existence of artificial autonomous entities that can interact with humans.*

Keywords: AI alignment, values, artificial autonomy, Value Alignment Problem, AI ethics

In this position paper we propose to embark in a systematic study of values and their relation with the autonomy and governance of artificial intelligent systems. The core question that motivates such effort is the *value alignment problem in AI*; that is, how to design artificial intelligent systems (AIS) that are aligned with human values. Our proposal is grounded, firstly, on the insight that autonomy in artificial intelligence systems (AIS) is the property that drives the most salient features of the positive and negative outcomes of AI research and development, and thus a topic that deserves a systematic study. Secondly, it is grounded on the insight that values may play a

*ORCID: 0000-0003-1317-2541; 000-0003-1283-8188

fundamental role in governing such autonomy, both to contend with the negative consequences of an artificial agency, as well as getting such agency to accrue value.

In this chapter we characterise four settings that illustrate the scope of the value alignment problem in AI and motivate the development of an AI-inspired theory of values through four settings: (i) the design of value aligned hybrid online systems, (ii) the modelling of values-driven behaviour, (iii)-an experimental setting for value engineering and (iv) the practical concerns of the governance of artificial agency. To establish our position, we outline the distinctive features of the four setting.

The organisation of this chapter is as follows: Section 2 contains a concise description of the value alignment problem and the underlying notions. The next four sections concern each of the settings just mentioned and Section 7 sums up the issues raised by the four settings and sketches an argument for an AI-inspired approach to alignment. In this position paper we propose to embark in a systematic study of values and their relation with the autonomy and governance of artificial intelligent systems. The core question that motivates such effort is the *value alignment problem in AI*; that is, how to design artificial intelligent systems (AIS) that are aligned with human values. Our proposal is grounded, firstly, on the insight that autonomy in artificial intelligence systems (AIS) is the property that drives the most salient features of the positive and negative outcomes of AI research and development, and thus a topic that deserves a systematic study. Secondly, it is grounded on the insight that values may play a fundamental role in governing such autonomy, both to contend with the negative consequences of an artificial agency, as well as getting such agency to accrue value.

In this chapter we characterise four settings that illustrate the scope of the value alignment problem in AI and motivate the development of an AI-inspired theory of values through four settings: (i) the design of value aligned hybrid online systems, (ii) the modelling of values-driven behaviour, (iii)-an experimental setting for value engineering and (iv) the practical concerns of the governance of artificial agency. To establish our position, we outline the distinctive features of the four setting.

The organisation of this chapter is as follows: Section 2 contains a concise description of the value alignment problem and the underlying notions. The next four sections concern each of the settings just mentioned and Section 7 sums up the issues raised by the four settings and sketches an argument for an AI-inspired approach to alignment.

1 Background

1.1 Motivation

The Value Alignment Problem

In this chapter we explore the problem of "alignment" in AI.[1] For that purpose we adopt the following definitions:

> *An AI system (AIS)* is a machine-based system that, for explicit or implicit objectives infers, from the input it receives, how to generate outputs such

[1] See Gabriel (2020)

as predictions, content, recommendations, or decisions that can influence physical or virtual environments. Different AI systems vary in their levels of autonomy and adaptiveness after deployment.[2]

The Value Alignment Problem in AI (VAP) is the design of Artificially Intelligent Systems (AIS) whose behaviour is *objectively* aligned with human values.[3]

S. Russell framed the VAP in the context of the debate on Artificial General Intelligence,[4] stressing the need to control AI autonomy as the distinctive feature of artificial intelligence systems where that risk originates, and suggesting the possibility of imbuing moral behaviour in AIS as a flexible and powerful way of governing that autonomy.[5]

Autonomy

The basic understanding of an autonomous AI system is that the AIS decides how to act in the world without the direct intervention of a human in *that decision*.

As summarised in Chapter 15 of this volume,[6] Moral Philosophy, Law, Political Science, Psychology and even AI have developed the notion of autonomy along different lines. In the context of AI alignment, that chapter discusses three ways of understanding artificial autonomy: one is the distinction between automated and autonomous with respect to military applications.[7] The second interpretation of autonomy is captured through the five well-known levels of autonomy for self-driving vehicles of the US National Highway Traffic Safety Administration.[8] This understanding is established in terms of a combination of the complexity of a task and the intervention of the human in the loop.[9] The third way of understanding is through another hierarchy of autonomy based on a combination of the complexity of the action (predefined delegation of one function vs. composition of several actions), the scope of the decision (goal specific, generic), and the involvement of humans in the support (off-line, crowdsourcing) or the use of an autonomous decision (to support the human's decision, no intervention).[10]

For VAP, we need an interpretation of the notion of autonomy as it relates to governance and values. What matters then are three main aspects: what type of action

[2] OECD (2024)

[3] Russell (2014) We relax S. Russell's phrasing of VAP's from a proof-theoretic alignment to an objective way of measuring to what extent the behaviour of an AIS is consistent with a specific set of values. We prefer *human* values (in Russell (2014)) over *beneficial* in Russell (2017).

[4] See Russell (2021)

[5] The full quote in Russell (2014) contextualises the motivation:"No one in the field is calling for regulation of basic research; given the potential benefits of AI for humanity, that seems both infeasible and misdirected. The right response seems to be to change the goals of the field itself; instead of pure intelligence, *we need to build intelligence that is provably aligned with human values. For practical reasons, we will need to solve the value alignment problem even for relatively unintelligent AI systems that operate in the human environment.* There is cause for optimism, if we understand that this issue is an intrinsic part of AI, much as containment is an intrinsic part of modern nuclear fusion research".

[6] Casanovas and Noriega (2025)

[7] See David et al. (2016), p.-4 and Cha. 15, Sec. 1, in this volume.

[8] In fact the NHTS labels six levels (0-5) of autonomy, but level 0 ("Momentary Driver Assistance") really doesn't count. See NHTSA (2018) and Chapter 15, Sec. 5.1 in this volume.

[9] See also Chap. 15, Sec. a5.1, in this volume.

[10] See Chap. 15, Sec. 5.2

is involved (goal directed, capability-specific, generic capability), what enables that action (principal-agent delegation, self-motivation) and whether or to what extent other stakeholders previous decisions or actions enable the decision and the action of the autonomous agent who acts.

These three aspects are needed to establish the basis for governance: (i) How to enable and modulate (promote, constrain) an autonomous agent's actions (discussed mostly in the value engineering cycle), (ii) how to allocate responsibility in practice (to credit benefit accrual or attribute blame) and (iii) how to improve value alignment in future autonomous actions.

Governance

In terms of VAP, governance has a dual function: To control risk and to foster value accretion in the operation of AIS, and in order to perform those functions on can draw on the standard governance devices.[11] One needs to incorporate those devices into AIS as *instruments* that promote value alignment.

As discussed in Setting 1, there are essentially four types of instruments:

1) *affordances* that enable the possibility of the AIS to execute an action,
2) *constraints* in the form norms, procedures and standards that affect the behaviour of individual agents or a collectivity of agents;
3) *information* that the autonomous system may incorporate into its decision-making;
4) *institutional agents* that interact with the AIS in order to provide added affordances, enforce governance constraints or provide inputs for the AIS decision-making.

Values

We adopt a rather standard interpretation of values that can be summarised in the following definition:

> "Values are beliefs linked inextricably to affect, refer to desirable goals that motivate action, transcend specific actions and situations, serve as standards or criteria, and are ordered by importance relative to one another (depending on context and "value owner") and the relative importance of multiple values guides action".[12]

An made operational with these six assumptions:

1) Values motivate and legitimise goals.
2) Values determine preferences between states of the world.
3) Values are contextual.
4) Values may be in conflict.
5) Values are hierarchic

The four settings show how one may interpret the way values are brought into the design of AIS and show that the resulting AIS they are value aligned.

[11] One needs to keep in mind what is distinctive of AIS, see Noriega and Casanovas (2025) for a discussion of AI risk and the role of values.

[12] From Schwartz (2012)

1.2 Running example

In order to keep our remarks in the following sections grounded, we propose to keep in mind two particular examples of AIS: (i) Self-driving vehicles, mostly cars trucks and buses of different types; and (ii) the transit coordination system that enables and governs the circulation of autonomous as well as non autonomous vehicles in a city.

To further illustrate some points we will occasionally mention other examples in the following sections.

2 *Setting 1:* Designing value-aligned hybrid online social coordination systems.

This is the VAP at its simplest, most straightforward version: to design one specific hybrid online social coordination system —with a particular purpose, specific context of application and specific stakeholders— that is meant to be aligned with an explicit list of values.

The distinctive features of this setting are (i) that we limit our discussion to a type of AIS which is a *restricted environment* —a hybrid online system that articulates the (social) interactions among artificial and natural autonomous agent, a "HOSS", for short— and (ii) that value alignment applies to the hybrid online *system as a whole*. The focus of attention in this setting is the design process of the HOSS itself, in particular on the methodological aspects of *engineering values into that HOSS*.

In addition to the traffic coordination system (that coordinates human and self-driven vehicles), this setting includes typical *online market-maker platforms* —like *Uber* or *Airbnb*— and *online multiagent-based management systems* like ambulatory health services in a large hospital.

While the core of this setting is the understanding of the design problem of engineering values into an AIS, the intent is, first, to *identify design heuristics and methodological guidelines* that may be applied to the engineering of values into online AIS and, second, to *characterise particular classes* of online AIS where one can build actual value-aligned AIS by applying specific heuristics and guidelines. As a side result, this setting aims to identify properties, heuristics and guidelines that can be applied to the engineering of values beyond HOSS. We will explore this point in our discussion of settings 2, 3 and 4.

2.1 Assumptions

The following provide a crisper description of the problem and serve to motivate and justify heuristics and methodological guidelines for the value engineering cycle of the next subsection. When these assumptions are refined they may characterise objectively value *alignable* classes of HOSS.

Hybrid online social coordination systems (HOSS).

As opposed to other settings, in this case we limit the value alignment problem to a closed-world environment where both natural and artificial autonomous agents can interact. The limitation has some advantages

(i) HOSS include "hybrid" systems where natural and artificial participants (agents) may interact.
(ii) It allows for a convenient differentiation between individual and collective value alignment, and in particular to focus on the alignment of the system as a whole.
(iii) It presumes all activity is online and mediated by the system interface. Thus, agent actions are, actually, messages that go through the system interface. Moreover, all relevant information (originating in agent activity or external events) becomes available to participating agents also through the system interface.
(iv) Since all interactions are online, the system determines what agent can enter it and, at any given time which individuals are active in the system at and what each can and cannot do.
(v) The system interface functions as the governance mechanism that enables affordances and enforces constraints.

Note that while these features do not apply to all AIS, it nevertheless captures an increasingly large class of actual online systems, as suggested by the examples above.

Situatedness.

We assume that we are dealing with a working system that fulfils some purpose for its users and that this system will be working online, thus the HOSS is situated in a concrete working environment. More specifically,

(i) The system, participants and working environment are distinct entities.
(ii) Participants' decision-making models are *opaque* to the system (participants decide on their own and the HOSS has no control on what and how they make these decisions, this is handled in Setting 2.).
(iii) The AIS is meant to be *fit for a concrete operational purpose* (speed up safe traffic; facilitate car rides).
(iv) *Compatibility with context of application.* If the HOSS is to be fit for an intended purpose it needs to be compatible with, for instance, *prevailing technological standards* (SDV features, communication platforms), *legal requirements* and procedures, (SUV certification, traffic regulations and enforcement), and *socio-economic context* conventions and practices (availability of certain types of SUV; traffic density, preferred routes).

Objective stance.

While in its original formulation, Russell required provability of value alignment, we relax that condition and only require "objectivity". These are the minimal commitments:

(i) There ought to be an observable way of determining whether the system is aligned or to what degree to each, to some or to a combination of the chosen explicit set of values).[13]

[13] In terms of the assessment of value alignment, design should take into account, both the alignment of the HOSS with respect to a given set of values that apply to the system as a whole (and are the result of some consensus among stakeholders), and the alignment of the HOSS with respect to the particular values of the different stakeholders.

(ii) Value alignment should be assessed from the specification or from the actual performance of the system.[14]

Stakeholders.

For design purposes one can assume that there are always three stakeholders classes (individuals or collectives) with their own values, interests and motivations.

(i) *Owner* who commissions the system, supports its construction and decides if and when it is deployed, updated, stays operational and decommissioned. Interested, for example in deploying a popular and practical traffic regulation system; or a market-maker who brings together riders and taxis and captures a large market share with adequate profitability and customer satisfaction.

(ii) *Builder* who is responsible for the technical aspects involved in the design, implementation and support of the working HOSS. The system is technically sound, resilient, reliable; responds to users needs and preferences; implements responsibly the objectives of the owner.[15]

(iii) *Users*: individuals —with properly established identity— who enter the system to satisfy their individual objectives, and have resources and entitlements that allow them to establish and fulfil commitments in the HOSS. There may be several classes of users with their distinctive capabilities, entitlements, needs and motivations (f.i, residents —who want fluid and silent traffic—, car manufacturers —reliable and growing market—, drivers, riders and car owners —safe, fluid traffic).

Engineering design assumptions.

As mentioned above, this setting is essentially a design engineering process whose specific assumptions are:

(i) It presumes three design commitments that draw inspiration from standard design methodologies:
   a. Value-sensitive. The system is meant to be imbued of values.[16]
   b. Full-cycle: from conception to decommission.
   c. Participatory: stakeholders are involved throughout the design cycle. In particular, stakeholders can reach consensus on value-specific design decisions.

(ii) Values can be imbued into the system through a *value engineering cycle* that consists of three main tasks: value selection, interpretation and alignment assessment discussed in Subsec. 2.2 and outlined in Fig. 1.[17].

[14] The objective stance can be refined further by assuming the HOSS is *state-based*. Namely (i) there is a finite set of variables that define the "state of the system"; (ii) the score of each of these variables is observable while the HOSS is active; and (iii) those scores change only when an attempted action becomes effective in the system or a relevant external event is recognised by the system. If we assume that the HOSS is state-based, then value alignment may be assessed as a function of the state of the HOSS and values can be promoted, protected or demoted by curtailing or enhancing the effects of agent actions and events (see next section).

[15] See, for instance, discussion of "conscientious design values" in Noriega et al. (2021).

[16] Value engineering should account for generic as well as context-specific values.

[17] See Noriega et al. (2021, 2023)

2.2 The value engineering cycle

Recall that the point of this setting is to build actual systems but with a distinctive character: they are, supposedly, imbued with values. This purpose is in line with the postulates of "values in design"[18] although, in this setting the point is to establish an understanding of the notion of value that can help govern collective interactions of self-motivated AIS so that the results are consistent with those values. Intuitively, in fact, the idea is to interpret values as a way of establishing "courses of action" where values postulate the main objectives of the system, the preferable means to reach them and the criteria for determining how good is the alignment at some point in the life-cycle of the HOSS.[19]

We propose to organise such design process as a five-task *value engineering cycle* described in Fig. 1. It starts with a primitive version of the HOSS that becomes enriched by identifying the values that stakeholders require from it, then these values are into means and ends that are programmed into the system in such a way that stakeholders may eventually claim that the working specification of the system is actually aligned with the values they have chosen.

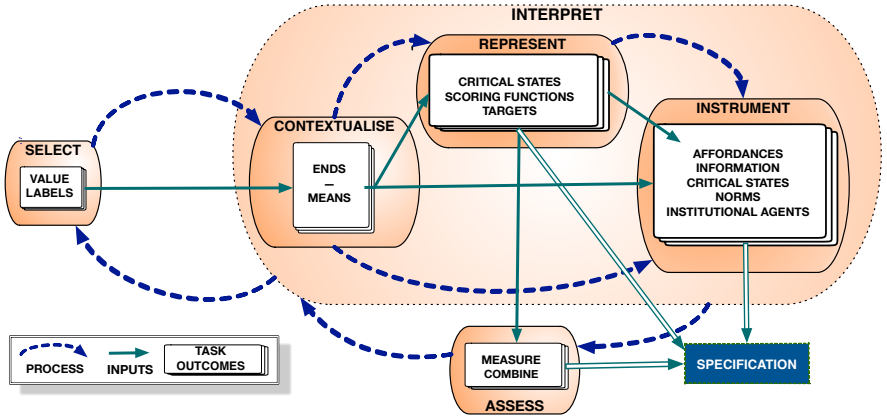

Fig. 1 Values are "engineered" in a HOSS through a series of tasks that translate an intuitive understanding of the relevant values into a set of specific ends and means that are implemented into the system so that its operation conforms with those values in an objective, measurable degree. (After Noriega et al. (2021))

T1: Value Selection

Stakeholders should reach a consensus on a list of labels that stand for values the system is meant to align with (safety, sustainability, traffic fluidity, residents' approval). These labels capture an intuitive shared understanding of the values the system should

[18] See Friedman et al. (2013)
[19] See Simon (1957, 1955) for a similar discussion around organisational goals.

be imbued with. Although their meaning needs to become precise and objective in the next three tasks.[20]

Note that, in addition to *context specific* values that determine the fitness of the HOSS to its intended purpose, labels need to reflect the preferences of stakeholders.[21]

In practice, there may be different heuristics for identifying these labels: Value taxonomies,[22] online value mining[23] and conventional preference elicitation set-ups (for instance, focus groups, customer surveys).

Once a set of value labels are settled upon, they need to be *interpreted* into objective ends that are motivated by the chosen values and a collection of devices that can be programmed into the HOSS in order to protect or promote the values that legitimise those ends. Value interpretation can be decomposed in three tasks:

T2: Value contextualisation

This task elucidates how the intuitive interpretation of values (safety, efficiency, fairness) can be turned into an *ordered list of ends* that are motivated by those value. This requires an agreement on a *collection of features* that became explicit in the previous task (accident types, numbers and severity; average fuel consumption; average speed in peak time on selected routes; cost and comfort of public transportation; ...). Some of these features will eventually be transformed into the precise parameters needed to represent values in T.2.

In addition, these *features* also serve to identify what actions that take place inside the HOSS, as well as well as external events, involve these features. This connection is needed for identifying a *list of means* that can be used to promote or protect each of the postulated ends and, therefore, need to be accounted for in the system specification (for example, traffic control sensors and actuators, incentives to car manufacturers for deployment of "civic" driving cars, subsidies for ride-sharing).

In sum, this task affects the HOSS specification by identifying (i) what is the relevant part of the physical world that needs to be captured into the HOSS (what actions should be enabled for users and what entities and parameters are involved in the activation and execution and potential governance of those actions); (ii) the sort of affordances and constraints that will be involved into the governance of the system; and (iii) the type of interfaces —and compatibility requirements— that should be implemented in the operating HOSS.

In this task, the point is not to reach a consensus on the details but to find common understanding on two main topics: (i) *Ends*: How each value should be observable in the system so that one can objectively claim that a given value is

[20] Although this is the usual starting task in software engineering, it may be worthwhile revisiting when consensus is hard to reach in subsequent tasks. In particular, this task —specially, when executed as the kick-off exercise— may result in the decommissioning of the system when consensus on labels and ordering is not reached.

[21] In most cases, different stakeholders have different value preferences. For example, the *owner* (city management) would prefer the traffic control system to be safe, sustainable, inexpensive and popular. *Users* like residents would prefer it to be say, safe and fluid, and car manufacturers would prefer it to be easy to comply with and advantageous to SDV over regular cars. And the *builder* would like it to be safe, efficient and resilient. Consensus, in this example would probably include most of these values in some fashion although some (e.g. "safety") most likely carry different meanings for different users.

[22] See Schwartz (1992); Rokeach (1973)

[23] For example, Liscio et al. (2021)

being supported or not; for example, *safety* of a transit management system may be measured in terms of accidents and traffic violations; *civility* in autonomous vehicles may be measured in terms of its norm-abidance disposition and with its track record on agreements and conflicts with other vehicles or drivers while in operation. And (ii) *means*: What actions can users take or, in general, what affordances and constraints should the system promote or constrain so that those ends can, in principled, be reached; for example, establish limited access zones with automated gate-keeping, require self-driving vehicles to have black-boxing with full disclosure, subsidize collective transportation.

The main heuristics have to do with the identification of goals and a subsequent goal decomposition process that produces a means-ends analysis of those main goals. The outcome should be a consensus on specific goals (motivated by the consensus values) to be reached with the system and with the means (identified by all stakeholders) that can be used to achieve these. Ideally, this consensus on ends should also bring a consensus on an ordering of those ends and at the same time some consensus about the most relevant means to implement.

T3: Value representation

This task determines how to value alignement becomes observable. Thus, it needs to bring into the specification of the HOSS those parameters that may stand for expected, desired and undesirable outcomes of the activity in the HOSS (traffic jams, rates of fatal accidents, traffic violations); as well as those elements that, based on those parameters, will support an objective assessment of the alignment of the system to the value or values in question.

There are two main approaches to providing observable and realistic indications of alignment. They are not incompatible.

1. *Critical states.* Values are represented as sets of situations, procedures or conditions that either ought to hold or ought not to ever hold. For instance, guarantee fluid traffic in peak hours and no traffic deadlock ever; SDV civility certification process; validation of automatic gate-keeping compliance.
2. *Performance indicators.* Values are associated with observable parameters that are updated while the system is in operation. For example, traffic saturation patterns in restricted access zones, number and types of accidents, residents' satisfaction.

The key heuristic is that one needs to achieve a consensus on the representation of each value and then take into account not only that consensus representation for a consensus alignment assessment, but also account for different value representations that suit the motivations and interests of the different stakeholders in the final alignment assessment.

T4: Value instrumentation

This task identifies the actual affordances and constraints that should be engineered into the HOSS specification in order to promote and protect the chosen values.

These are some heuristics for tasks 2, 3 and 4:

- Making goals and instruments explicit usually requires an update of the ontology, affordances and constraints of the HOSS.
- Generally speaking, one can choose among five devices: (i)Critical states of the system (that have to be avoided or reached), (ii) affordances that participants can take advantage of, (iii) norms, procedures and other constraints that regulate interactions, (iv) information that participants may use, and (v) autonomous agent participants whose decision models are defined by the HOSS.
- These tasks may be approached with classical AI devices. Namely, since values motivate goals (as we mentioned in Sec. 2) one can use conventional AI goal-decomposition and the associated means-ends analysis.
- If one adopts the *state-based assumption* ends take the form of a utility function on state parameters. Means are those devices that promote actions that lead to better scores and discourage or prevent worse scores (see Setting 3).

T5: Value alignment assessment

One needs to specify into the HOSS how to tell whether it is aligned with the chosen values. There are essentially three conditions to meet: One is to tell if the agreed upon ends of the system are effectively met, another is to see if the agreed upon means are adequate for reaching those ends, and finally, how satisfied are the stakeholders with the system.

The first one, *effectiveness*, depends on the objective stance and in its simplest form is based on the fact that the system captures the *consensus representation of the values*, thus validating that critical states are properly implemented and performance indicators are met to a reasonable degree (e.g.the system prevent circulation collapse, signalling didn't go over budget, fuel consumption is within acceptable limits and serious accidents are rare).

The second, *adequacy*, can be understood as a cost-benefit analysis of alternative instrumentation of the consensus values (the cost of "digitalising" the city vs conventional traffic lights and police; dedicated lanes to SDV and free parking). And the third amounts on a combination of the assessment of each stakeholder of the effectiveness and adequacy of the system but with respect to the values of that stakeholder (Are residents happy with the changes? What about car manufacturers? Is the major likely to be reelected?).

These three forms conditions can be expressed as functions and depending on the representation of values and the way the cost-benefit trade-offs measured, one can turn value assessment as some sort of multiobjective decision-making process.

2.3 Comments

The challenge is to design actual systems that support a collective activity that brings together natural and artificial autonomous agents in such a way that the activity is accomplished objectively in accordance with an explicit set of values.[24]

[24]This setting draws on Noriega et al. (2021, 2022, 2023)

Although this purpose guides the "values in design" agenda[25] and was the subject of attention in the early work on organisational design,[26] in this chapter we explore two AIS-specific angles: The characterisation of a class of AIS that can be value-aligned by design, and a value engineering cycle —with concomitant meta-heuristics— that applies to this class and, *mutatis mutandis*, applies also to other types of AIS. Notice also that, as suggested in Setting 3, the value engineering cycle can be developed to account for alternative understandings of the role of values within a given AIS and be applied in a systematic way.

From a methodological perspective, our proposal distinguishes three classes of design stakeholders (that in various guises are also present in the other settings). We also acknowledge the need to reach consensus among them, while also taking into account their different and sometimes conflicting values. This consensus needs to be achieved in all stages of the value engineering cycle and during the complete life-time of the online system.

The assumptions we used to characterised HOSS and the simple meta-heuristics we discussed can be refined to characterise a large class of value-aligned systems like the ones in Setting 3. In particular, by refining the **dialogical** stance and the **observability** stance, one can characterise a class of HOSS that can be objectively value aligned, as discussed in Setting 3. This refinement is made precise for the class of *online institutions*.[27]

A different type of extensions of the HOSS class that ought to be studied are those online AIS that can be "provably" aligned.

Although this setting has been presented as a process, of engineering values into a system, it also provides elements for the assessment of the alignment of an exiting system with respect to other values. This leads to at least a pair of related technological products: certification of value alignment and the development of add-ons that govern an existing system towards a desired value alignment.

3 *Setting 2:* Modelling values-driven behaviour.

This setting serves to explore how to design autonomous agents who act in accordance with some explicit values. In particular the modelling of the decision-making process of an AIS as defined in OECD (2024).

While in setting 1 we were concerned mostly with the governance of the aggregate outcomes of collective interactions, in this case we turn our attention to the *self-governance* of individual autonomous artificial agents.

The approach in this setting is interdisciplinary: How different disciplines contribute to the understanding of the values-driven process and how these contributions may be used to engineer values into a situated autonomous AIS. In this section we will simply point out the obvious insights of some disciplines as an invitation for further analysis.

[25] See for example, Friedman et al. (2013); Davis and Nathan (2015); van den Hoven et al. (2015); van de Poel (2020)
[26] See Simon (1957)
[27] See Noriega et al. (2023).

3.1 Assumptions

Stakeholders

In this particular setting, the focus is on one particular stakeholder: The autonomous agent whose values are the ones that motivate its behaviour. Note that the autonomous agent's behaviour may need to acknowledge those values associated to the environment where it is situated. Note also that from a design perspective, there still are the usual three classes of stakeholder: Builder, owner and user (as in Setting 1) with their own list and interpretation of values and, as discussed in Setting 4, each with an associated liability.

Situatedness

As it is the case in other settings, the autonomous entity in question has a behaviour that responds to some design purpose (however flexible, vague or unpredictable it might be) and work properly in a given operating context.

In this light, one issue is to establish what are the aspects of the operating environment that can be perceived and those that can be affected. A second issue is to decide how those percepts are incorporated in the decision model: For example, if percepts become beliefs or fact in the decision model, and how. And third, what type of action the agent decision triggers since this is determined by the type of autonomy, the impact of its actions, and the type of self-governance vs compliance with external governance that needs to be engineered in the agent's interface with is operating context.

Decision models

The point of this setting is to study whatever takes place "inside" an autonomous agent, between perception and action and in particular how values are involved in what determines that the agents' behaviour is value aligned. Abusing language, we call that "black box" between perception and action the agent's *decision model* and the focus of this setting is to make such black box transparent.

Such decision models can be of any sort, from strictly hardwired value-aligned default behaviour, learning and adaptative self-governance and deliberative decision-making in various forms. Thus, strictly speaking, an action that an agent takes —think of SDV or the guardian agent— is not necessarily a deliberate moral choice, not even a decision. The point of this setting is precisely the exploration of the multiple aspects that might be involved in the agent's response, how these aspects can be modelled and engineered, and, prominently, how values are embedded in such modelling and engineering.

Value engineering

This setting favours a conceptual analysis of value alignment and may profit from the engineering and practical aspects that may be more preponderant in the other settings.

One focus of research here is whether or when behaviour is driven by one or several values, whether it is context-specific, or whether it is persistent or adaptable

to external conditions. Likewise, in this setting it is worth distinguishing how and when the agent's behaviour is aligned with its own individual needs and preferences, and how and when it is well adapted to the context where it is being active.

Mental states and other behaviour driving constructs.

The way values-driven behaviour is approached may provide motivation and solution to some of the most contentious aspects of autonomy that are explored in Setting 4. Namely, what agent models support, validate or discard the practical consequences of artificial moral agency. What types of mental states could be modelled in order to address whether or to what extent malevolence is an issue and benevolence its counterpart in artificial agency. Likewise what should be an essential aspect of individual decision model to render it accountable.

3.2 Multidisciplinary approach

We propose to approach this problem space from four vantage points.

A cognitive view of values-driven behaviour.

Explain value-driven behaviour in terms of several cognitive features, one of which are values. Fig. 2 describes a rather general "mental model" of a cognitive agent who chooses to perform an action after processing new input. The model and how values modulate the relevance of potential actions may involve typical deliberative constructs —like knowledge and beliefs— together with other constructs that capture persistent features of the individual agent —like motivation and personality— others that capture more transient features —like the prevalent needs of the agent and the resources and entitlements to execute an action that becomes available— as well as others that capture social aspects of the decision behaviour —like awareness of other agents' values. This figure also suggests the possibility of modelling the choice of individual actions and the role of those cognitive constructs in the configuration and updating of plans.[28]

Logics-inspired models:

A wide variety of deliberative forms of values-driven behaviour, like reasoning with values, reasoning about values, arguing about values or using values in argumentation, can draw inspiration and be formalised through logics-based systems and techniques. For instance, values-driven decision-making con be modelled directly with the arsenal of formalisms of deductive and inductive reasoning (for instance conventional and ad-hoc modal, preference, possibilistic logics); as well as combinations of these (like context logics) to mirror distinctive concomitant forms of reasoning in a cognitive architectures.

[28] In addition to the modelling of deliberative values-driven behaviour, cognitive science as well as neurology and behavioural economics may inspire alternative modelling. For example, in Rangel et al. (2008), the authors propose three types of value-based decision-making: "Instinctive" predetermined behaviour, adaptive learning-based and deliberative.

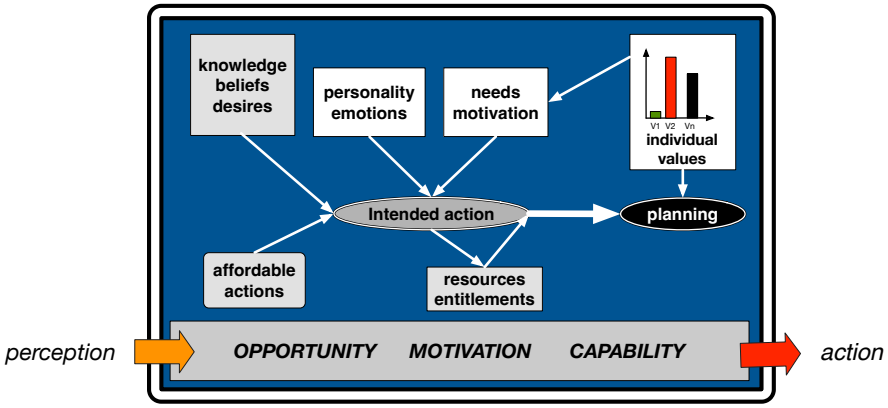

Fig. 2 An agent decision making model. The actions that an autonomous agent takes depend on the state of the world it can perceive and a decision to choose an available action in accordance to the agent's cognitive mind-frame, values and plans.

Likewise, one can draw from dialectical and argument-based reasoning formalisms to model, for instance, the assimilation of perceived data, update beliefs, negotiate priorities and devise individual and cooperative action plans.

Other approaches:

Several formalisms associated with rational choice can be conveniently adapted to values-driven behaviour. Those intuitions provide grounds for several conceptual and formal developments that can be readily used in the modelling of values-driven behaviour. Game theory, operations research and, in particular, multi-objective optimisation are good examples of conceptual frameworks. Related techniques like organisational design, business strategic planning and corporate goal alignment also provide metaphors and tools for modelling VDB.

Moreover, the vast corpus of metaphors, concepts, principles, heuristics and examples of economics and political science, beyond the particular contribution around rational choice, are a rich source of inspiration and tools for modelling value-driven behaviour.

Finally, most of the original motivation of value alignment in AI is a mirror of the problem of value alignment in philosophy. Thus the corpus of ethical metatheory, normative ethics and applied ethics are of fundamental value for this setting.

4 *Setting 3:* Sandboxes for policy design

While in Setting 1 we were mostly concerned about the methodological aspects of value engineering, and in Setting 2 with the modelling of self-governed agents, in this case we discuss how to design an artefact that provides an experimental platform to test alternative approaches to value engineering and the implementation of realistic instances of values-driven agents. Policy design sandboxes are such artifact.

A sandbox is a simulation workbench built on top of what is essentially a hybrid online social coordination system. The fact that policy-making is a *values-driven*

process and simulation involves values-driven stakeholders, makes the design of workbenches a particular case of value aligned AIS design. What sets this setting apart from Setting 1 is that the problem of value engineering needs to be addressed at different levels of abstraction, and that the adequacy of such engineering can be tested experimentally using the sandbox.

The sandbox serves two main purposes. The first is "epistemic": to explore the effects of a *policy intervention* —a specific sets of policy measures— in a given *policy domain*. For example in city traffic management: deploying street sensors and traffic tracking systems, granting circulation privileges to SDVs. The second is "rhetorical": as a device to reach consensus among stakeholders (say authorities, car manufacturers and residents for example) on what is the *best* policy intervention to deploy.

4.1 Assumptions

Stakeholders

In this setting, one should distinguish three classes of stakeholders:

1. *Policy owners* are the individuals or groups that represent public entities, firms and citizens who are responsible for the choice and deployment of a policy intervention. For example, city government, Ministry of Transportation, car manufacturers, residents). Note that, in terms of Setting 1, these are in fact the *main users of the workbench*.
2. *Policy-designers* are (i) the individuals who develop the simulation model, test it and decide to make the workbench available to the policy owners: that is, for example, city planners, city mayor, local transit authority, associations of residents, car manufactures; as well as (ii) the engineers who design, implement and test the simulator.
3. Finally, *policy-subjects* these are individuals and groups whose activity within the policy domain is affected by an intervention. They can be actual persons who "participate" in simulation runs or simulated agents (who attempt a policy enabled action only when the agent is motivated, capable and has the opportunity).

Notice that the different stakeholders hold different values and that the type of values that govern stakeholders decisions are different in each of the classes. Notice also, that consensus among stakeholders needs to be achieved at three levels: First, on the engineering of values in the simulator. Second, in the settings and validation of the experimental process and the recommendation of policy interventions to policy owners. Finally, policy owners need to reach consensus on the quality, relevance and political feasibility of a policy to be deployed.

Simulation model

In order to serve the *epistemic function*, the simulation model ought to be *realistic* and *reliable*: provide good relevant predictions in order to assess the effects of interventions. It is built around an agent-based domain model with three main components:These three components capture the persistent features of the simulation

assumptions while transient features are captured in simulation scenarios and the actual policy interventions.[29]

1. A *physical model* that captures the content and dynamics of the relevant "natural" affordances and constraints of the domain environment that enable the instrumentation and assessment of policy interventions: actions of policy subjects, events and the entities that are required to enable actions and observe their effects (purchasing an SDV; traffic signalling, grid capacity and loads; fuel costs; $CO-2$ emissions; average travel times, etc.).
2. An *institutional model* that captures the "artificial" constraints that articulate agent interactions: norms, regulations and social practices (like local taxes, circulation code, enforcement capabilities, elective sustainability practices, required SDV safety features).
3. A collection of *agent models* of those stakeholders that are active within that policy domain with their own different and often conflictive needs and preferences, and reflect in a realistic way the evolution of the policy domain (f.i., human drivers with different driving behaviour and self-driving vehicles of different kinds and manufacturers). [30]

Values in policy design and simulation.

In this setting, values are present in several guises:

- Policy-subjects decision-making.
- Definition of a policy objectives.
- Choice of policy means.
- Assessment of the success of a policy intervention.

Their engineering process is the same of Setting 1. In particular, the assessment of the success of a policy intervention is a combination of three value-alignment functions mentioned there:

- *Effectiveness*: the degree to which policy objectives are achieved
- *Adequacy*: trade-offs between alternative interventions (cost-benefit analysis)
- *Acceptability*: the degree with which policy subjects find the intervention aligned with their own needs and preferences.

The rhetorical use of a policy-design sandbox

Sandboxes, in general, provide a sort of *ex-ante* "evidence-based" assessment that is specially useful to for the exploration of unusual or atypical approaches. Value-driven sandboxes have the added advantage of being "informative" in a particular way. They capture and reveal the interplay of the motivations of individual policy subjects and they capture and reveal the assumptions that support the choices of goals and instruments, especially the evaluation of the potential success of an intervention. In so doing, value-driven sandboxes provide arguments in favour of particular policy interventions.

[29] These simulation models are an instance of the HOSS in Setting 1 and in particular online institutions Noriega et al. (2023).

[30] The behaviour of these individuals can be modelled as instances of Setting 2 agents who choose to enact a given policy-enabled action only when they are capable and motivated and given the opportunity to attempt it.

The simulation sandbox is, in its basic form, a tool for a systematic exploration of policy interventions. As described in Fig. 3 this exploration consists of testing versions of a particular policy intervention, re-testing with alternative scenarios and fine-tuning the simulation model in the process. This systematic process provides evidence-based arguments that facilitate consensus among policy designers and support their deployment recommendations to policy owners.

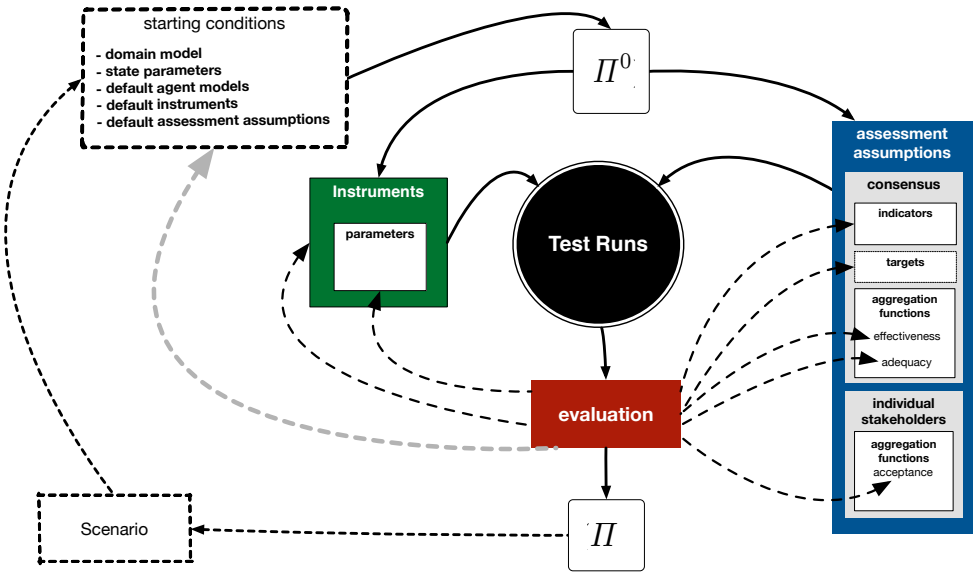

Fig. 3 The experimental cycle for testing and choosing policy interventions (adapted from Noriega et al. (in press)).

4.2 Comments

Notice that the simulation model of a policy sandbox in Setting 3 is in fact a "value-alignable" HOSS. Sandboxes provide concrete examples where the different tasks of the value engineering cycle needs to be instantiated.[31]

Note also that these sandboxes support a more specific version of the *objective stance* in Setting 1. Namely, it is *state-based*: (i) One can define explicitly what constitutes the *state of the world at any given time* (the score of a finite number of policy indicators), (ii) The state changes only when either an attempted action a policy subject, or an external event is recognised by the system. This two extra assumptions make it possible to assess value alignment as a function of the state of the system and conceive value instrumentation as either the modulation of policy-subject actions or conditions that guarantee critical states (regiment their achievement or make them

[31] Some ideas for this setting were developed by the authors previously in Noriega and Plaza (2022, 2024); Noriega et al. (in press).

inviable). Consequently, Setting 3 provides a rather general tool for the testing and validation of value alignment of the class of AIS that share those assumptions. By the same token, it supports a large variety of practical applications of this setting.

Third, from an application perspectives, these sandboxes have an epistemic advantage over other forms of modelling by capturing behavioural and values dispositions of policy subjects in an ostensible way. For this same reason, and for the systematic process of value-driven policy analysis and fine-tuning —with the explicit involvement of those stakeholders that decide the deployment of a policy— sandboxes provide a rhetorical advantage for negotiation not readily available otherwise.

5 *Setting 4:* Governing artificial agency in the wild

The focus in this setting is on artificial systems whose autonomous actions have an actual effect in the real world. The approach, in this case, is to explore those real-world aspects involved in the actual governance of their activity. In this setting we will use a (software) financial assistant as an example.

We can outline the problem with three questions: (i) What are the actual risks and potential benefits specific to artificial agency? (ii) What conventional notions, devices and procedures are at hand to deal with the unwanted effects? and (iii) What are their shortcomings? We propose to study the three questions from a conceptual, methodological and empirical perspective, focusing on the governance of unwanted and problematic aspects but extracting general guidelines for accruing benefits.

While in Setting 1 we were concerned with the problem of engineering online environments where AIS could be corralled into value-aligned behaviour, and in Setting 3 we focus in the way AIS can become self-governed, in this setting we take a wider perspective and propose to explore artificial agency *ex-ante* —what can be done to prevent harm and to foster benefits— and *ex-post*— how to deal with potentially ill-behaved proactive AIS.[32]

As a meta-heuristic, we propose to address this exploration through the combination of risk analysis and legal perspectives. In this section we sill focus on risk analysis and let legal remarks be dealt in Chapter 15 of this volume.[33]

5.1 Assumptions

On stakeholders.

For this setting we will refer to the autonomous artificial system as an *agent*, and refer to the usual three design stakeholders classes (owner, builder and user). However, from both a risk-management and a legal perspective, the definition of these stakeholder classes needs to capture a precise understanding of how the notions of harm, impact and responsibility apply to each stakeholder. These nuances are essential so that risk can be sensibly decomposed and redress procedures can be properly implemented. Thus, in this setting, the *builder* is, as usual, the person, team, supplier who designs, deploys and supports the system itself and is therefore responsible for the correct

[32] Although in this section we ostensibly contend with the negative aspects of autonomy, most of our remarks can be translated into heuristics to proactively derive benefits associated with artificial autonomy.
[33] See Noriega and Casanovas (2025); Casanovas and Noriega (2025).

operation of the artificial agent. In Setting 4, instead of "owner" we prefer to call *supplier-manufacturer* the person or company who commissions the design of the agent and makes it available (sells, licences, authorises) to an *owner-user* of the agent, so that the artificial agent can act in the wild. Note, however, that the term "user" —now "owner-user"– requires clarification because, depending on the way the supplier-manufacturer transfers the use of the agent to the owner-user, there are significant differences in terms of responsibility of builder, supplier-manufacturer and owner-user.

In particular, the notions of supplier-manufacturer and owner-user need a subtler analysis because in order to allocate responsibilities, we need to distinguish two sharpy different forms of *agency*:

(i) *Delegated agency* where the actual owner-user of the agent is the *principal* who enables and entitles an (artificial) agent to act in its behalf, and therefore the principal is responsible for the effects of the agent's actions; and

(ii) *Un-delegated agency* where an *autonomous artificial agent* takes actions that affect the real world but responds to no principal since it might not be explicitly instructed by its owner-user, nor the supplier-manufacturer to take those consequential actions. For the purpose of redressing the effects of a casualty, in this case, it becomes essential to elucidate the corresponding responsibility of builder, supplier-manufacturer and owner-user in order to render each dully accountable.

Finally, because the autonomous artificial agents in question may interact with other AIS, one needs to identify these *third parties* as another class of stakeholders who may be involved in casualties.

Situatedness

Artificial autonomous agents are situated in a particular working environment and ought to be compatible with the existing conditions. As discussed in Setting 1, the artificial agent needs to be compatible with technological, legal and social conventions and requirements in order to work properly.

5.2 Analysis of AIS risk

Risk is usually understood as the unwanted consequences of a contingent event. More technically, risk is the value that results from a combination of hazard (an unwanted outcome), impact and likelihood. Risk can be managed by taking some measures to control these components.[34]

In our example, the risks of using a personal financial assistant to organise one's travel is, for example, to pay too much for a trip that is not what we really wanted. Likewise a patient guardian agent may put a patient's health at risk if it conducts a poor analysis of drug interactions and fails to rise an alarm, or it may risk rising insurance costs or delaying a patient's treatment by requiring unnecessary clinical tests.

Like we suggested above, risk management —and, respectively, value accretion– may de addressed through an *ex-ante* analysis of potential risk and the design of those measures that can reduce it, and an *ex-post* analysis of what to do when an unwanted

[34] Recall that, in this section, the discussion is centred in the negative aspects of fortuitous events but most remarks can be adapted to the beneficial outcomes of a contingent event.

contingent event materialises (see Fig. 4). This separation and the corresponding analysis originates in the insurance industry but applies broadly to risk-management practices in general.[35]

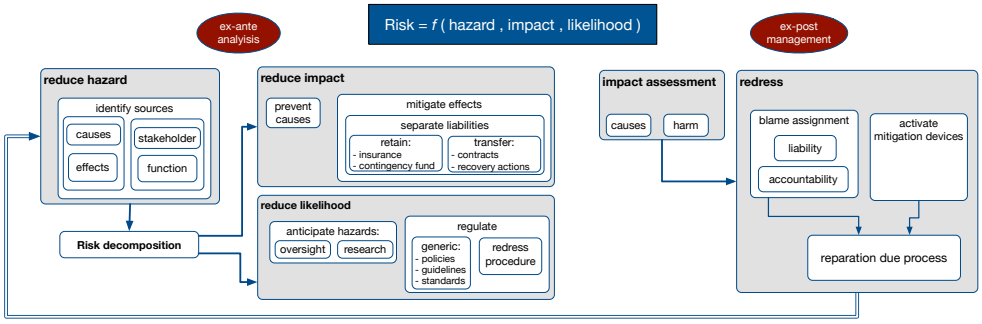

Fig. 4 The risk-management process includes an *ex-ante* analysis process where the three components of a given risk (hazard, impact and likelihood) are analysed in order to devise ways to reduce their overall effect. There is a second —*ex-post*— risk-management process to deal with the consequences of an actual occurrences of the risk (a "casualty").

5.2.1 Ex-ante analysis

The purpose is to reduce risk by finding ways to address systematically its three components: hazard, impact and likelihood.

Track 1: Separate risks.

The purpose is to reduce harm by identifying the main sources (or causes) of the original risk, determining the specific harm associated with each of these so as to subsequently address the specific impact and likelihood of each of these sources.

This decomposition or separation is achieved from two viewpoints: First, by identifying those hazards that are specific to the *functionality* of the AIS and, second, by identifying how these different hazards are associated with the stakeholders involved in the design, manufacturing, ownership and use of the AIS.

For example, a personal travel assistant would be capable of, say, booking lodging. Main sources of harms for the principal include overspending and opportunity cost. Each of the two can be decomposed further and, for instance, set a credit limit policy to avoid catastrophic expenses without compromising opportunity cost. However, the artificial agent may perform actions with unintended consequential effects for which these preventive measures are not enough —say book an non-existent tourist apartment that is nevertheless listed in the lodge-booking market-maker— and in this case one should then be able to assert whether these casualties are due to misuse (due to lack of information about the sytem and its operation or to negligence of any of the

[35] A similar discussion is described in Noriega and Casanovas (2025), since AI-risk is essentially not different from other types of risk.

stakeholders involved, including third parties) or to malfunction of the artificial agent or the booking service, and then allocate responsibility accordingly.

Track 2: Reduce impact.

For each risk that results from the previous analysis the idea is to take preventive measures that reduce the exposure and damage to stakeholders. That is, restrict who may be adversely affected by a casualty an the extent to which those who are exposed have to bear the adverse effects of a casualty.

The default procedure is to avoid a potential harm altogether by removing the original sources of risk, specially when risks have already been separated. Another way of avoiding casualties in conventional risk-management practice is to limit exposure, which in this case is achieved by curtailing agent autonomy, mostly in terms of agent functionality and access to third parties.

Since potential harm can be avoided only to some extent, one can still reduce impact by mitigating the effects of unwanted outcomes when a casualty takes place. The key reduction is achieved by identifying the *liability* of each stakeholder. That is, the responsibility of that stakeholder for imposing some sort of harm. Once liability is established —and depending on the particular risk, stakeholder, and cost— the full impact of a casualty will be divided among all liable stakeholders. This way, each stakeholder is burdened only with its own liability. In practice, each stakeholder may decide how much of that liability to absorb and how to transfer the rest to a third party modulo some kind of contract. Typically, the economic liability that stakeholder retains is usually mitigated with an insurance policy or, in general, coverage by some contingency fund. Other types of liability are addressed through analogous devices to attenuate moral damage (e.g., emergency task-force and recovery protocols) and the corresponding provisos and resources (PR campaigns).[36]

When, in spite of these preventive and mitigation measures, an unwanted harmful event impacts stakeholders an ex-post *redress procedure* needs to be enacted to compensate damage, prevent unwanted side effects and devise better harm prevention and impact mitigation measures, as we discuss below.

Track 3: Reduce likelihood.

This process involves essentially two elements (for each separate risk identified in Track 1 above): First, setting up means to anticipate the emergence of new risks. Essentially oversight mechanisms and research actions that provide early warnings on new unwanted outcomes and emerging potential hazards. Second, the deployment of preventive regulatory measures to avoid unwanted outcomes and support redress processes; that is, regulation of hazardous functionalities, for determining stakeholders' liability (specially supplier-manufacturer/user-owner), and to render stakeholders accountable; as well as policies, operational guidelines, best practices, industry standards and the like.

[36] Note that we are using "liability" in a wider sense: the stakeholder obligation may not necessarily be legal and the harm to account for may be not only monetary, see below 5.2.2. Recall that in this section we presume that our remarks extend *mutatis-mutandis* to the positive counterparts of risk, harm and liability.

5.2.2 Ex-post management

Once a harmful unwanted event takes place, stakeholders need to contend with damage. This involves an impact assessment to identify actual damage and revise ex-ante assumptions and provisos, and the enactment of a redress process to identify liabilities, compensate losses and restore damage.

Track 4: Impact assessment.

When a risk materialises, *mitigation actions* and available safeguards become active immediately reduce actual impact and to avoid cascading effects. For instance, freeze credit lines, cash-in insurance coverage or disable certain functionalities in the agent or in its operating environment. Then two lines of analysis start: the first one is the elucidation of the causes and circumstances of the event. The second is a sound estimation of the costs (not only economical) associated it. This analysis is performed for the casualty as a whole and for each of the stakeholders.

This two-pronged impact analysis is essential for establishing the facts and conditions for the activation of a redressing process. Moreover, the analysis of each casualty also provides an opportunity to improve the current risk management plan (i.e., the whole ex-ante process).

Track 5: Redress process.

This is meant to help stakeholders recover from the harmful effects of the materialisation of the contingent event. It can be organised as two complementary tasks: The first is to establish stakeholders responsibilities and entitlements. The second is to embark into a process of *reparation*.

While conventional practices are quite appropriate for the impact assessment (Track 4) —and we will not comment further on these— reparation of AI induced harm requires special attention to the issue of agent autonomy in order to identify liabilities, render stakeholders accountable and compensate harm.

First a word on liability: In conventional terms, like in automotive insurance, (i) the stakeholder who suffers an accident is entitled to a compensation granted by the underwriting conditions established in its policy and (ii) the insurance company is liable to cover the compensation costs. The *underwriting* conditions establish that the corresponding insurance fee ought to be in good standing at the time of the accident, whether there is any deductible down-payment before the insurance takes care of reparation costs, what type of incidents are not covered and what is the top compensation amount. If conditions hold, the insurance company is "liable" in a narrow sense: it has the legal obligation to pay the corresponding amount for underwritten damages. For our purposes, we need a wider term. First, we need to take into account not only financial costs but include "moral harm" of different sorts. In addition, to render the responsible stakeholder accountable for the moral harm, we also need to consider other not only legal mechanisms to "allocate blame". In fact we need to establish the *responsibility* of the materialisation of the harmful event. In our case, this is usually a problem of "many hands responsibility" and, in principle, there are conventional means to address it. However in the case of AI-induced harm the elusive aspect is the degree of autonomy delegation involved in artificial agency that we mentioned above.

Presuming one can establish liability, there should be a legal framework in place to back the process of rendering the responsible stakeholders accountable and bring them to repair damage. Note that for liability to be properly allocated, such legal framework would need to establish the conventions to determine the mode of involvement of the responsible stakeholders not only as victims but also in the cause of the mishap.

The framework that backs the reparation would include not only established procedures for establishing liability and render parties accountable, but also the relevant policies, guidelines, supervisory bodies and international agreements to make it operational.

Note that what makes this reparation process different in artificial agency is, as mentioned above, autonomy. If and when an autonomous artificial agent is deemed liable, the procedure ought to determine if the casualty is due to a malfunction of the agent or a misuse, and whether this failure is due to negligence or incompetence of the artificial agent and, eventually, if malevolence can be ostensibly predicated, and, reparation should proceed accordingly. The key issue, then, is how to render this autonomous agent accountable. This is not always clear.

If autonomy is explicitly delegated, accountability should belong to the principal but, as suggested above, who the principal is may not be always clear. Moreover, delegation and responsibility may be blurred among the capabilities and entitlements of several supplier-manufacturer and the owner-user stakeholders involved in the casualty to some degree.

5.3 Comments

Our discussion in this setting has focussed on the unwanted outcomes of artificial agency. However, it should be clear that a similar analysis of value-accretion can be systematically performed and thus turn value alignment and artificial agency into a powerful and beneficial technology.

The aspect that distinguishes this setting from the previous three is the fact that we assume that an artificial agent can have effects on the world and those effects may be harmful. Conventional risk management and legal devices and means are well suited to address many of the unwanted outcomes. Nevertheless, artificial agency has a peculiarity that escapes these conventional resources: When an agent causes moral harm (and depending on the degree of autonomy of that agent), it may be extremely difficult to establish liability and therefore properly render accountable such agents to compensate the harm they produce.

However, depending on the formalization involved in the modelling of individual behaviour, provability of alignment –for a single agent– may be easier to achieve in this setting, thus rendering support to the allocation of responsibility.

It is beyond the scope of this section to discuss how "un-delegated autonomy" can be formally and effectively addressed with legal means. If artificial agents are granted some form of legal personhood, the problem of rendering them accountable remains even when some form of bonding is attached effectively to their operation.

Legislation and other forms of regulation may be necessary to allocate responsibility in the mechanisms of possession and activation of an artificial agent, however enforcement of such legislation is doubtful.

Along these lines, settings 1 and 2 provide some useful governance devices, specially when agents have properly instrumented delegated autonomy. From a methodological perspective, this setting suggest an additional task for the value engineering cycle of Setting 1: the impact assessment of a deployed AIS and how that impact can me managed and become legally bound.

The deep sources of risk in autonomous agency are tightly coupled with three factors of the embedding of AI technology in products and services: the opacity of the specific technologies, the speed of release and adoption of technology, products and services and the business models of the preponderant players. The legal framework mentioned above needs to address these factors in earnest.

6 A proposal for the study of alignment in AI

6.1 Potential contribution of the four settings for the study of alignment in AI

Setting 1: Design of value aligned hybrid online systems

Setting 1 is a restricted version of the value alignment problem (VAP).

In broad terms, the challenge in this setting is to explore the function of values in the design of AIS and more specifically how values determine courses of action within the system. This setting explores the ideas of "values in design" but applied to the specific case of autonomous artificial systems and their collective governance, and through the methodological approach of a value engineering cycle.

The leading questions in this setting are: What are values? How can they be used to govern the collective activity of autonomous entities? and What does it mean that an AIS is aligned with a value?

What distinguishes this setting form other restricted versions of VAP is the fact that the focus is on the alignment of a *situated system as a whole* and not of a generic AIS. This restriction suggests two additional assumptions: (i) *The dialogical stance*: all interactions are mediated by the system that can enforce the entitlement of a participant for the use of affordances, the compliance with constraints and the effects that actions have in the system. (ii) *The objective stance*: alignment can be assessed through the state of the system and the embedded functionalities.

One direction for future research is de refinement of the assumptions we use in this setting to characterise actual classes of AIS that are value aligned.[37]

This setting can also be extended in order to encompass the value governance of individual AIS (Setting 2) and other AIS, not only HOSS; and for identifying the scope of liable autonomy is the use of artificial agency in practice (Setting 4).

Finally, this setting can be used to design closed environments that encapsulate AIS. In other words, designing online regulated environment where artificial agency is no longer in the wild.

[37] See Noriega et al. (2023).

Setting 2: Modelling values-driven behaviour

The focus is on the individual's use of values, to understand how values govern individual behaviour and how to assess alignment.

The main purpose of this setting is, first, to show the advantages of a multidisciplinary approach to study the notions of autonomy, the social aspects of individual decision-making and the concerns about the situatedness of individual value-driven behaviour. And, second, to study the positive and the negative aspects of individual autonomy, its governance and the role of values. In particular, this setting is conducive to the study of alignment in proof-theoretic and other formal terms.

This setting provides case studies, heuristics and guidelines for the design of value-driven autonomous agents to be used in settings 1 and 3 but, more significantly, objective and concrete use-cases, heuristics, guidelines and standards for the design and assessment of the artificial agents in Setting 4. In particular, it provides formal elements to determine potential and actual impact and liability of agents in the wild and thus the elements for risk-management and due process that concern the owner-manufacturer of an agent, as well as the third parties involved in the activity (support, interaction and governance) of these agents.

Setting 3: Policy sandboxing

From a conceptual point of view, sandboxes provide a convenient AIS framework for the notion of value and its implementation and from a normative perspective, the distinct roles values play in the sandbox provide a variety of perspectives for the process of value engineering.

From a methodological perspective, sandboxes provide a technology for a systematic testing of alternative instantiation of the value engineering cycle. Thus, the identification of value engineering heuristics of different sorts, both for individual AIS (Setting 2) and for value-aligned collective AIS interaction (Setting 1).[38]

In addition, since the simulator of a sandbox is in fact a value-aligned AIS, this setting actually provides concrete practical examples of value aligned HOSS that can be extended, adapted or inspire other actual HOSS.

From an application perspective, value-driven sandboxes can be used as a powerful simulation workbench to study social phenomena in multiple domains. Although in terms of values, when values-driven sandboxes are used as a tool for policy design, they serve as a flexible negotiation instrument because they provide evidence-based support to proposals and an agreement setting for design stakeholders and policy owners. In fact, this type of value-aligned platforms support the selection, deployment and follow-up of actual policies that may have substantial social and economic value.

This setting provides a tool to support impact assessment of a potential or deployed AIS (not only policy-making) and thus guide the risk analysis as well as the risk management process outlined in Setting 4.

[38] In other words, the policy selection process in Fig. 3 can be used in Setting 1 to guide the design process of the online system.

Setting 4: Agency in the wild

Setting 4 addresses the type of practical concerns that ought to be taken into account when autonomous AI systems are deployed.

While the discussion of this setting was focussed on the harmful consequences of artificial agency, there is a dual notion of value accretion that may be addressed roughly along the same legal and risk management lines discussed there.

Perhaps the most significant conceptual problem that can be explored in this setting is that of responsibility as it concerns the different stakeholders involved with artificial agency; and, concomitantly, the notions of value-accretion and blame attribution.

This setting also makes prominent the discussion about the philosophical aspects of moral agency in AIS and the legal requirements to contend with is risks.

This setting also establishes a conceptual and methodological space to study to what extent conventional governance approaches are well adapted to risk control and value accretion when they are applied to the activity of autonomous agents; and how they need to be judiciously to be better the adapted to contend with autonomy of AIS.

6.2 Towards an AI-inspired approach to alignment

The four settings outlined in this chapter are meant to support a research programme based on the following 4 tenets:[39]

1. We propose to extend the scope of classical value theories to encompass the activity of individual artificial autonomous agents as well as collective action situations where artificial and natural autonomous agents interact. We propose the notion of artificial agency as the key construct and a problem domain outlined by the interplay of the notions of autonomy, governance and, of course, values.

2. We claim it would be AI-inspired because we propose to build on top of AI constructs, methodologies and tools and to embark in such study like early AI did on the study of cognition: with a wide perspective on the leading intuitions, with interdisciplinarity, and a set of paradigmatic problems to explore a variety of associated notions.

3. For the development of this AI-inspired theory of values we propose to take the value alignment problem in AI as a guiding light.

4. The four settings we discuss in this paper might be developed as the type of paradigmatic problems for an AI-inspired theory of values. If nothing else, they provide enough hints of the type of conceptual developments, engineering methodologies and constructs, as well as practical applications and concerns associated with the value alignment problem.

Acknowledgements. Authors wish to acknowledge the contributions and fruitful discussions with Pompeu Casanovas, José Antonio Donaire, Mark d'Inverno, Julian Padget, Manel Poch and Harko Verhagen. Research for his paper is supported by CSIC's (Bilateral Collaboration Initiative i-LINK-TEC) project DESAFIA2030 BILTC22005; EU (Horizon-EIC-2021-Pathfinderchallenges-01) Project VALAWAI

[39] Some ideas for this setting were presented by the authors previously in Noriega and Plaza (2024).

101070930; and the EU (NextGenerationEU/PRTR program) and the Spanish (MCIN/AEI-10.13039–501100011033 program) project VAE TED2021-131295B-C31.

References

Casanovas P, Noriega P (2025) Governance of Artificial Agency and AI Value Chains: A Few Remarks on Autonomy from a Legal and Ethical Approach., Springer

David R, Nielsen P, Allard J, et al (2016) Final Report of the Defense Science Board Summer Study on Autonomy. Publicly-Releasable Version. URL https://apps.dtic.mil/sti/citations/AD1017790, [Online] Retrieved Sep 2025

Davis J, Nathan LP (2015) Value sensitive design: applications, adaptations, and critiques. In: Handbook of ethics, values, and technological design: Sources, theory, values and application domains. Springer, p 11–40

Friedman B, Kahn PH, Borning A, et al (2013) Value Sensitive Design and Information Systems, Springer Netherlands, Dordrecht, pp 55–95. https://doi.org/10.1007/978-94-007-7844-3_4, URL https://doi.org/10.1007/978-94-007-7844-3_4

Gabriel I (2020) Artificial intelligence, values, and alignment. Minds and Machines 30(3):411–437. https://doi.org/10.1007/s11023-020-09539-2, URL https://doi.org/10.1007/s11023-020-09539-2

van den Hoven J, Vermaas P, Van de Poel I (2015) Handbook of ethics, values and technological design. Springer

Liscio E, van der Meer M, Siebert LC, et al (2021) Axies: Identifying and evaluating context-specific values. In: Proceedings of the 20th international conference on autonomous agents and MultiAgent systems, International Foundation for Autonomous Agents and Multiagent Systems, pp 799–808

NHTSA (2018) Levels of Automation. URL https://www.nhtsa.gov/vehicle-safety/automated-vehicles-safety, [Online] Retrieved Sep 2025

Noriega P, Casanovas P (2025) From the Pascal Wager to Value Engineering: A Glance at AI Risks and How to Address Them. In: Value Engineering in Artificial Intelligence: Second International Workshop, VALE 2024, Santiago de Compostela, Spain, October 19–24, 2024, Revised Selected Papers. Springer-Verlag, Berlin, Heidelberg, p 257–275, https://doi.org/10.1007/978-3-031-85463-7_16, URL https://doi.org/10.1007/978-3-031-85463-7_16

Noriega P, Plaza E (2022) The use of agent-based simulation of public policy design to study the value alignment problem. In: Casanovas P, de Koker L, et al (eds) Proceedings of Selected Papers of the Workshop on Artificial Intelligence Governance Ethics and Law (AIGEL 2022), CEUR Workshop Proceedings, vol 3531. CEUR-WS.org, (Ret. Oct 2024), pp 130–139, URL https://ceur-ws.org/Vol-3531/SPaper_10.pdf

Noriega P, Plaza E (2024) On Autonomy, Governance, and Values: An AGV Approach to Value Engineering. In: Osman N, Steels L (eds) Value Engineering in Artificial Intelligence. Springer Nature Switzerland, Cham, pp 165–179, https://doi.org/https://link.springer.com/chapter/10.1007/978-3-031-58202-8_10

Noriega P, Verhagen H, Padget JA, et al (2021) Ethical online AI systems through conscientious design. IEEE Internet Comput 25(6):58–64. https://doi.org/10.1109/MIC.2021.3098324, URL https://doi.org/10.1109/MIC.2021.3098324

Noriega P, Verhagen H, Padget J, et al (2022) Design Heuristics for Ethical Online Institutions. In: Ajmeri N, Morris Martin A, Savarimuthu BTR (eds) Coordination, Organizations, Institutions, Norms, and Ethics for Governance of Multi-Agent Systems XV. Springer International Publishing, Cham, pp 213–230

Noriega P, Verhagen H, Padget J, et al (2023) Addressing the Value Alignment Problem Through Online Institutions. In: Fornara N, Cheriyan J, Mertzani A (eds) Coordination, Organizations, Institutions, Norms, and Ethics for Governance of Multi-Agent Systems XVI. Springer Nature Switzerland, Cham, pp 77–94

Noriega P, Poch M, Jose A D (in press) Sandboxing sustainable tourism. In: Bush A (ed) Sustainability and Artificial Intelligence - Reshaping Tomorrow, SpringerNature, vol 14520. Springer, pp 165–179

OECD (2024) Explanatory memorandum on the updated OECD definition of an AI system. 8, https://doi.org/https://doi.org/https://doi.org/10.1787/623da898-en, URL https://www.oecd-ilibrary.org/content/paper/623da898-en

van de Poel I (2020) Embedding values in artificial intelligence (AI) systems. Minds and Machines 30(3):385–409

Rangel A, Camerer C, Montague PR (2008) A framework for studying the neurobiology of value-based decision making. Nature Reviews Neuroscience 9(7):545–556. https://doi.org/10.1038/nrn2357, URL https://doi.org/10.1038/nrn2357

Rokeach M (1973) The nature of human values. Free press

Russell S (2014) Of Myths and Moonshine. A conversation with Jaron Lanier, 14-11-14. The Edge, URL https://www.edge.org/conversation/the-myth-of-ai#26015, [Online] Retrieved Oct 2024

Russell S (2017) Provably beneficial artificial intelligence. The Next Step: Exponential Life, BBVA-Open Mind

Russell S (2021) Living with artificial intelligence. URL https://www.bbc.co.uk/programmes/b00729d9/episodes/downloads

Schwartz S (1992) Universals in the content and structure of values: Theoretical advances and empirical tests in 20 countries. Advances in experimental social psychology 25:1–65

Schwartz SH (2012) An overview of the Schwartz theory of basic values. Online readings in Psychology and Culture 2(1):11

Simon HA (1955) A behavioural model of rational choice. The Quarterly Journal of Economics 69(1):99–118

Simon HA (1957) Models of man; social and rational. Wiley