From the Pascal Wager to Value Engineering: A Glance at AI Risks and How to Address Them

Pablo Noriega[1 [0000−0003−1317−2541]] and
Pompeu Casanovas[1,2,3 [0000−0002−0980−2371]]

[1] Artificial IIIA-CSIC
Campus UAB, Barcelona, 08193 Spain
pablo@iiia.csic.es
[2] UAB Institute of Law and Technology, (IIIA-CSIC Associated Unit). Campus UAB, Barcelona, 08193 Spain
[3] La Trobe Law School
Melbourne, 610101 Australia
pompeu.casanovas@iiia.csic.es

Abstract. There is widespread awareness of the potential undesirable consequences of AI. In this paper we explore the grounds of such concerns from the perspective of risk management. We propose a decomposition of AI risk into three sorts of risk (inertial, disruptive, fundamental) that can be approached in different but complementary ways. From this differentiation we advocate for Value Engineering as a pertinent approach to address fundamental AI risk.

Keywords: AI risk, engineering values in AI artefacts, autonomy, governance, risk-management

1 Motivation

This paper is an invitation to equanimity. We look into the risks associated to AI, delineate a general strategy for dealing with them and focus on the role that value alignment can play in taking care of some of those those risks.

There are three main biases in our discussion: First, that the assessment of risk is based on exposure, impact and likelihood. Second, that conventional risk-management approaches and simple common sense are quite useful in the assessment of AI risks and devising adequate strategies to address them. And third, that we focus our attention on *AI artefacts* —for the specific type of risks they pose—, and not on all potential risks associated with AI as a scientific discipline, nor those associated with the social phenomena emerging from human interaction with AI systems.[4]

The argument is structured around four claims: (i) The existential threat of AI is a latent risk that underlies more imminent AI risks, which can

[4] We use the expression "AI artefact" to stand for "artificial intelligent system"; that is, a machine based system that exhibits some sort of autonomy, in the sense defined in [39] and discussed below in Sec. 7.

be approached through conventional risk management strategies. (ii) By looking at the historical adoption of AI as a proxy for AI risk, one can decompose AI risk in three distinct categories: *inertial, disruptive* and *fundamental*. These types correspond to the way that, for forecasting purposes, time series are typically decomposed (trend, intervention and residual). (iii) Each of these three types of AI risk can be addressed with a judicious adaptation of conventional risk management practices. However (iv), fundamental AI risk, which emerges from the inherent autonomy and adaptability of AI artefacts, requires a more circumspect approach. And finally, (v) a prudent strategy to contend with fundamental AI risk is to use ethics and value engineering in particular as means to tame autonomy.[5]

2 From extinction threat to catastrophic risk

According to Stephen Hawking:
"Success in creating AI would be the biggest event in human history. Unfortunately, it might also be the last ..." [6]

Stuart Russell is more precise:
"*Given our current lack of understanding of how to control AGI systems and to ensure with absolute certainty that they remain safe and beneficial to humans, achieving AGI would present potential catastrophic risks to humanity, up to and including human extinction."* [7]

Figure 1 describes Russell's view of AI risk. It postulates that: (i) Harm caused by AI may be huge, but it is very unlikely although not impossible; hence (ii) today (t_0) risk is small ($0 < \epsilon$), and (iii) this situation remains rather stable until a future time, (t_σ), when a *singularity* takes place, namely Artificial General Intelligence (AGI) is achieved and takes control over human affairs.[8]

[5] There is a lively ongoing discussion about the interplay between risks, autonomy, values and AI that this paper addresses; for example, [32,40,22,4]. Such interplay is also addressed explicitly in some salient official documents, like [49,28,38] and is at the core of the EU AI Act [15].

[6] The Independent, May 4, 2014. https://www.independent.co.uk/news/science/stephen-hawking-transcendence-looks-at-the-implications-of-artificial-intelligence-but-are-we-taking-ai-seriously-enough-9313474.html (Retrieved Oct 2024) [23]

[7] Stuart Russell. Testimony to the Subcommittee on Privacy, Technology and Law of the US Senate Committee on the Judiciary. Accessible in the webpage of the 25.07.2023 hearing: https://www.judiciary.senate.gov/committee-activity/hearings/oversight-of-ai-principles-for-regulation).

[8] According to S. Ulam, the notion of a technological singularity —the moment (in a distant future) when the evolution of technology surpasses human capabilities— was probably first articulated by J. von Neumann [52]. Shortly after, Turing phrased it in AI terms: "*It seems probable that once the machine thinking method had started, it would not take long to outstrip our feeble powers... At some stage therefore we should have to expect the machines to take control*" [51]. More recently, Vinge, Kurtzweil, Bostrom and others popularised it in connection to chiliastic views like "transhumanism" and "longterminism" [11,3].

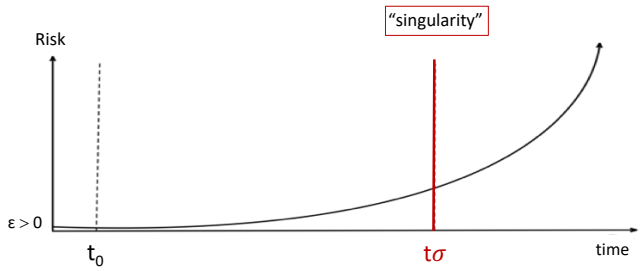

Fig. 1. The AI Singularity according to S. Russell: Currently (t_0), the risk of human extinction due to AI is negligible and will remain so until a future time (t_σ) when AGI is achieved and risk becomes catastrophic.

> While Hawkins' assertion articulates a threat, Russell's qualifies it as a latent risk. The advantage of phrasing Hawking's dictum in Russell's terms is that it makes the threat concrete: the extinction of humanity, because of AI, is a future event with an extremely large cost but a very small likelihood.
>
> In spite of the fundamental differences. Russell's phrasing echoes Pascal's Wager as a risk-mitigation approach: "Since we cannot have proof of the existence of God, the rational choice is to live a virtuous life" (Pascal, Blaise, "Infinite-Nothing", §233, in Pascal's Pensées [43]). [9]
>
> Russell's claim provides little guidance to elucidate the likelihood of reaching the singularity but, as Pascal puts it, one can bet on identifying other not latent AI risks and devise ways to manage them. To elucidate what those risks are we can rely on previous experience. Rather than merely contemplating extinction, one can look at how the adoption of AI has evolved over the years and try to identify features that suggest what impact AI has had, what types of hazard AI rises, and identify

[9] The key paragraph reads: "'God is, or He is not.' But to which side shall we incline? Reason can decide nothing here. There is an infinite chaos which separated us. A game is being played at the extremity of this infinite distance where heads or tails will turn up... Which will you choose then? Let us see. Since you must choose, let us see which interests you least. You have two things to lose, the true and the good; and two things to stake, your reason and your will, your knowledge and your happiness; and your nature has two things to shun, error and misery. Your reason is no more shocked in choosing one rather than the other, since you must of necessity choose... But your happiness? Let us weigh the gain and the loss in wagering that God is... If you gain, you gain all; if you lose, you lose nothing. Wager, then, without hesitation that He is."

strategies to anticipate and contend with AI risk. In order to guide that exploration towards an understanding of AI risks and how one can deal with them, we propose to rely on conventional risk-management heuristics. We will actually identify three different classes of AI risk which can be approached through different risk management strategies.

3 Addressing AI risk with a standard risk-management approach

We understand risk as the expected cost of an adverse but uncertain event. The purpose of risk-management is to reduce those expected costs. This reduction can be achieved with a combination of actions —the *risk-management process*— summarised in Fig. 2.

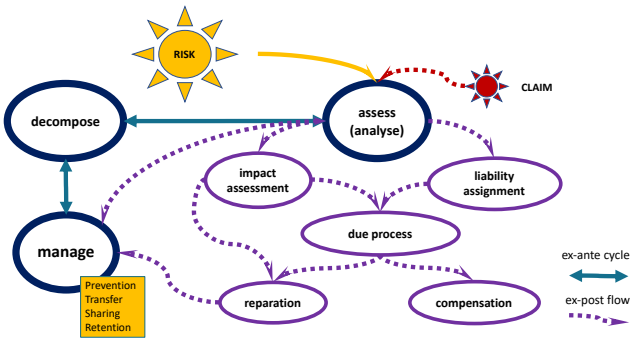

Fig. 2. The risk-management process. It includes an ex-ante cycle where specific risks are identified and analysed in order to define those actions triggered by a an actual occurrence of that risk (a "claim").

The process involves the identification and analysis of relevant risks and the specification of those actions that should take place if and when the risk materialises.

A risk is usually analysed through the assessment of five elements: (i) a *hazard* —the achievement of AGI; a wrong AI-based diagnosis—; (ii) *exposure* where and how and whom would be affected if the hazard actually takes place —human race; a patient, a pathologist and a hospital— (iii) *impact* the unwanted consequences of the materialisation of the risk —AI takes over human affairs; mistreatment of the patient— (iv) *likelihood* of the materialisation of the risk —unlikely but not impossible; some measurement of the effectiveness of post-diagnosis treatment— and

(v) *liability* a measure of the adverse consequences if and when the hazard materialises —extinction of civilisation; economic and moral damage compensation of malpractice claims.
Since the purpose of risk management is to reduce impact, one would try to distinguish what elements are involved in a *claim* (a presumable materialisation of the hazard, by analogy of insurance terminology) and deal with each separately in order to reduce the overall impact (AGI landmarks; wrong decision model, bad data, equipment malfunction, inadequate medical protocol, and so on).
Each of those fragmented risks should be managed on its own and ultimately put in place the mechanisms to mitigate the actual impacts of its occurrence (as suggested f.i. by Bengio et.al [4]). In the malpractice example the mitigation mechanism might be (i) revise current practice (main steps of the relevant protocol including the AI diagnosis decision-making), (ii) improve oversight, (iii) establish a procedure to deal with malpractice claims, and (iv) buy insurance to cover legitimate malpractice claims.
Once this ex-ante analysis-decomposing-management cycle is ready, the materialisation of a risk (a "claim") triggers a flow of actions that determine its impact, attributes blame, takes measures to compensate the adverse impact of the claim and using this experience improves the overall risk management process.
As we shall see below, a circumspect adaptation of this conventional risk management process can be applied to AI related risks.

4 A 3-fold decomposition of AI risk

Acknowledging that AI risk is correlated with the use of AI, we propose to estimate AI impact (and harm) from historical market information. That is, we can use the time series of the AI market value to identify different sorts of risk that are associated with AI, their sources and impact; and find ways to manage each of them.
Customarily, time series (like the one in Fig. 3) can be analysed, for the purpose of forecasting, as a combination of four components: (i) a

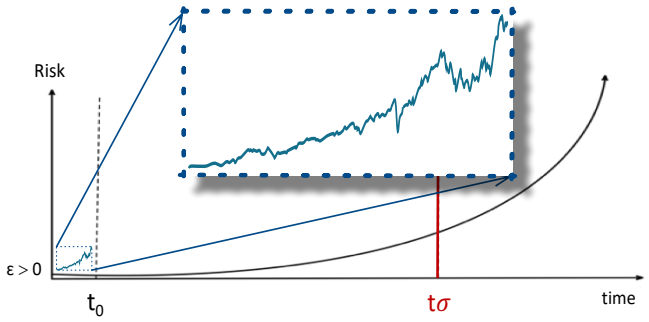

Fig. 3. Future risk can be estimated from historical data.

cyclic component that describes periodic and seasonal influences, (ii) a trend component that describes the long-term orientation of the series, (iii) "interventions", that alter the profile of the series with respect to other components and (iv) the core stochastic (residual) process. Forecast is based on the evolution of each of these components.

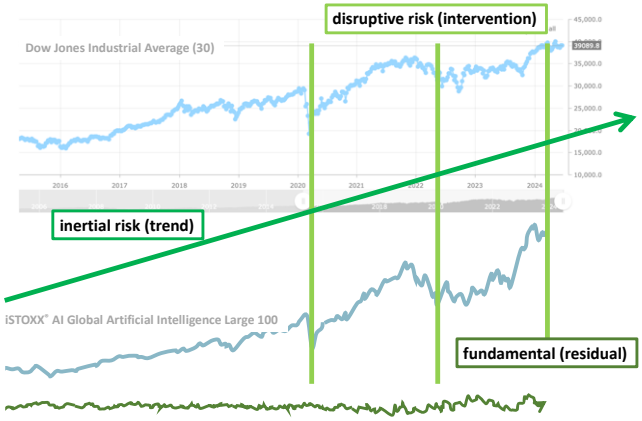

Fig. 4. A decomposition of stock-market prices into three components (trend, interventions and residual) and, by analogy, a decomposition of AI risk as *inertial*, *disruptive* and *fundamental*). The top graph shows the evolution of the Dow Jones index; the bottom one, the evolution for the same period of an index of one hundred large investors in AI (from: iSTOXX AI Global AI Large 100).

Based on the two time series in Fig. 4, we propose to project AI risks onto those time-series components in order to elucidate different sources or types of risk and devise differentiated management strategies.

Fig. 4 plots two stock-market time series from June 2014 to June 2024: the Dow Jones Industrial Average (30), on top, and the iSTOXX® AI Global Artificial Intelligence Large 100 index (https://stoxx.com/index/ixagal1p/). The top series shows rather clearly the three components we mentioned: an overall ascending *trend* (that absorbs a cyclic time-series component, which is irrelevant for our purposes), two prominent *interventions* (COVID and Fed intervention on interest rates linked with the Ukrainian war), and the underlying *residual* (stochastic) component. The bottom series shows that the segment of the stock market that is most influenced by AI has a patently similar profile. Thus, in the segment of AI companies one can also identify three components: the core stochastic component of the time-series would correspond to the risk that is directly linked with AI: the *fundamental* AI risk. The trend component corresponds to an *inertial* AI risk, which "pushes" the fundamental risk over time. And, finally there is a *disruptive* AI risk that originates from external, unanticipated events (the interventions) that resonate with the

fundamental risk and produce significant changes in both fundamental and inertial risks.[10]
Although all AI risks share some common features and basic risk management heuristics apply to all, each of these three types of AI risk has distinguishing features that are amenable not only to further risk differentiation but to a specific risk-management approach.

5 Addressing inertial AI risk (tendency)

From the above, we use the label *Inertial AI risk* to capture the underlying trend of the time series and its cyclic (periodic and seasonal) components. Intuitively, inertial risk reflects the historical evolution of risk, independent of the constantly evolving *fundamental risk* and independent of disruptive, unanticipated events coming from exogenous forces or radical AI innovations.
One can identify four main sources of inertial AI risk:

1. *Increase of IT power and online activity*, as enablers of AI innovation, demand and productivity.
2. *Underlying market inertia.* AI investment, development and use are affected by the overall business activity. Hence, to capture the influence of that general activity —that is largely independent of AI proper— one can use market indicators (market trend, cyclic components and interventions), that are extrinsic to AI but may still reflect forces that impinge on AI.
3. *Short term provisos* that influence amount, concentration and sources of investment in AI R+D+T —like regional, national and industry-specific policies, programs, incentives and regulations. Although, in the the short term they are disruptive for AI to some extent, they are also an inertial component by affecting direction, development and adoption of AI. Consequently, they are especially significant for the design of medium and long-term AI risk management.
4. *Maturity of AI.* Not only as a scientific discipline but also in terms of ready-use technologies, professional expertise and market development.

The combination of the previous elements determines long term evolution in the adoption of AI, and therefore in its positive and negative impacts. These elements suggest a cautious, *vigilant* risk-management approach. In practice this amounts to decomposing inertial risk and build on proven risk-management practices to design *ad-hoc* mitigation devices for each of those risk decompositions. For instance, (i) monitor evolution of AI impact through AI observatories, (ii) identify disruptive features in the

[10] Note, that for the two interventions the AI index has a steeper slope and wider variation. Although this difference can be explained in part because of the smaller number of firms in the iSTOXX index, another explanation is the fact that large companies that invest in the development and use of AI—like ALPHABET; AMD; META, NVIDIA and Siemens— have attracted significantly more capital in this later period (see Fig. 5).

demand and capitalisation of enabler technologies (e.g., emergence of social networks), (iii) include AI-specific considerations in policies that involve provisos on education and R&D, oversight organisms, and international agreements on these aspects (e.g., NSF and EU projects); and (iv) design AI-specific provisos to articulate some type of long-term responsible AI policy.

6 Addressing Disruptive AI risk (interventions)

As suggested by the analogy with time series interventions, disruptive AI risks come from unanticipated innovations, events or circumstances that have a profound effect on the types of AI artefacts that become available, the expectations and the actual use and adoption of AI. Hence affecting the preexisting risk profile.

Interventions can be seen as a sort of crisis. They produce a Thom-like singularity ([50]) that alters the magnitude and volatility of the AI market and all its subsidiary indicators; hence its *liability*. But they also trigger strong adaptive reactions in demand, supply, policy and social perception.

There are two illustrative historical examples of disruptive AI interventions. The first one is the so-called "AI winter" in the sixties: a substantial decrease in AI investment and development spawned by unfulfilled expectations of the field and a suspension of soft R&D funds. The second is the so-called "Fifth Generation Computer Revolution", in the early seventies, when the convergence of new computer architectures, the consolidation of knowledge-based systems and the expected adoption of AI in manufacturing and strategic decision-making brought in a "spring" of resources, motivation and expectations [31,17,24]. We are arguably in the midst of one. In addition to diffused reactions to Hawking-like concerns, two ostensible arguments back this claim: the distinct alteration of stock prices of three large AI players (Apple, Microsoft and Nvidia) in recent years (Fig. 5), and the dramatic increase in public interest sparked by ChatGPT captured in the OECD's observatory of AI-related news (Fig. 6).[11]

Fig. 5. Evolution of stock prices over the last ten years of Apple, Microsoft and Nvidia. From Apple's *stocks app* (retreived 14.10.24).

[11] Some voices claim that the current AI disruption is bringing the singularity closer. As close as the next decade [4]. We claim that such urgency needs to be reined in with equanimity.

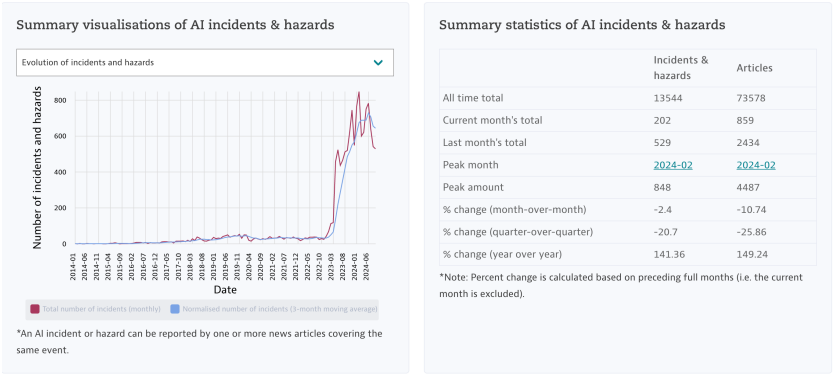

Fig. 6. OECD's automated monitor of AI news and hazards from public sources. From https://oecd.ai/en/incidents (retreived 14.10.24)

Even a crude description of the current crisis helps illustrate the key dynamics of disruptive AI risk one should address:

1. The convergence of fundamental and inertial risk components
   (a) A fundamental risk component from AI innovations that had been evolving over several years (essentially ML, NLP and MAS), and
   (b) An accelerated inertial risk from:
      i. the sophistication of AI enablers (GPU, "AI chips", cloud computing) that is being driven by the colossal requirements of AI processing.
      ii. The availability of massive digital content (text, images and knowledge);
      iii. Added thrust from the adoption of AI enabled applications.
2. Landmark success cases (from *Jeopardy and Alpha-Go* to *ChatGPT*) that stimulated important advances in different AI technologies and gave public visibility to AI.
3. An unprecedented market impact and speed of adoption.
4. Inordinate concentration of knowledge, capital and capital investment in very few dominant AI firms.
5. A swift reaction of stakeholders (academia, industry, authorities, press and public) stimulated by risk aversion, ambition and lack of information.

In light of the speed and force of the current disruption, it seems advisable to embark in a *reactive* risk-management strategy with two aims in sight: on one hand, to identify specific potential harms and elucidate what are the main factors that are playing out risk in the current disruption; and, on the other hand, identify stakeholders and their interests, in order to articulate adequate institutional mechanisms to manage the associated risks, and to address liability and accountability in particular.

Hence, a prudent risk-management strategy would suggest the following reactive actions:

- Identify potential impacts, triggering events and conditions.
- Adapt successful risk-management strategies and mechanisms from other risk-prone technologies and professional practices (health, energy, finance, genetics).
- Dilute risk concentration through incentives for start-ups and commit public investment in research and innovation on the disruptive topics.[12]
- Adequate legal procedures to support impact assessment and attribution of liability on AI-based claims, in order to guarantee accountability and foster risk mitigation.
- Analyse the role of AI in systemic risks (finance, health, energy, environment, defence, security) and take relevant protective measures.
- Develop new institutional frameworks to design and implement risk control policies.

7 Addressing fundamental AI risk

Fundamental AI risk is the one that derives directly from the deployment of AI systems and is not included in the previous two sorts. This is the risk that is intrinsic to AI and therefore difficult to characterise *a priori*. Nevertheless, one can still advance a preventive risk-management approach by trying to elucidate where the risk originates and analyse the forces at play. A sensible heuristic to find that source is to identify what makes AI different from other disciplines, or better yet, what are the most salient features that distinguish AI systems from other. We find that the following characterisation ([39]) is a good starting point:

"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 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".

This OECD definition carries three important implications. The first one is that from an AI risk-management perspective, the key aspect to address is *autonomy*. That is, the capability —engineered into AI systems— of making decisions that affect the world without direct intervention of humans [10]. The second one is that fundamental AI risk resides in AI artefacts. It resides not in the potential impact of *AI as a scientific discipline* concerned with understanding intelligence in the wide sense, nor is it about the *social phenomena* that result from the interaction between humans and artificial autonomous artefacts. The third one leads to preventive outlook: in the context of risk-management, one can understand

[12] This device reduces AI's negative impact and amplifies potential benefits by enlarging the set of stakeholders (those who possess knowledge, resources and competence to identify and contend with emerging risks and opportunities) and fostering a social appropriation of value form these innovations. As a side effect it discourages premature release of technology.

autonomy along two well-established notions of autonomy and deal with each separately. Namely, (i) Autonomy as a relationship of delegation between a *principal and an agent*, in the traditional legal and economics sense; and (ii) autonomy of a individual in the form of *moral agency*, in the philosophical and cognitive sense.

Autonomy as delegation. This notion of autonomy presumes two parties: a *principal* and an *agent*. The principal holds some capabilities, entitlements or authority that enables it to perform certain actions whose consequences the principal is liable for. Delegation, in this case, is a procedure through which some of those capabilities, entitlements or authority are passed onto an agent who will now be enabled to execute some actions on behalf of the principal.[13]

The key issue to keep in mind in delegated autonomy is that delegation of autonomy concerns not only the capability of "acting in representation of", but it also involves an actual allocation of *liability*.[14]

This type of delegation of autonomy between principal and agent has a long tradition and has been the source of several devices to support risk management tasks. Three examples serve to illustrate this form of delegation: (i) A car manufacturer delegates onto a car dealer the capability of selling cars, provide maintenance and acknowledge when a certain failure is covered by the manufacturer's guarantee. (ii) Highway police has the authority, empowered by administrative public laws, to assess traffic violations and to enforce norms to prevent life-threatening situations. (iii) A pathologist may delegate the task of screening tissue samples to an AI-based diagnostics system to simplify triage (identify positive cases and discard the rest).

In rough terms, the car dealer will be liable for the costs of poor repairs, not the manufacturer; and a policeman will be liable for any harm that is the consequence of a misinterpretation of its entitlement to use force, i.e. its abuse of power. Example (iii), illustrates how, in medical practice and other similar domains, risk-management is used to decompose risk in order to put appropriate risk mitigation devices in place and attribute different liabilities to the different stakeholders involved in the harm resulting behaviour. Namely, the pathologist is entitled to delegate the decision to label a sample "positive" because such delegation is part of the approved medical protocol and because the hospital has allowed

[13] From a legal perspective, delegation must be differentiated from *legal empowerment*. "Delegation" presupposes an already established power, which is delegated to another physical (individual) or moral person (an institution or organisation). "Empowerment" refers to the constitutive act of attributing to a person or an organization certain normative capacities, i.e. "powers", without the need for prior delegation.

[14] In our context, "liability" refers to the legal responsibility, to the economic cost attached to a harmful behaviour, and also to the associated social and moral damages. The issue of liability and rendering liable parties accountable has received attention not only in law but also in economics, see for instance [13,14,26]. It is worth noting too that, from a legal point of view, there is a well-known tradition linking accountability and autonomy with causation and liability.

the use of the specific AI system in this protocol. The hospital, in turn allows the use of the particular AI system because, in addition to the endorsement of the use of an automated diagnostics in the medical protocol, the supplier of the specific AIS provides some guarantees that the systems works properly. Similarly, the supplier signs a contract with the hospital that reflects its own confidence on the proficiency of the AI system, which is eventually based on proper testing and sound science and engineering. The point is that, in spite of this chain of entitlements, the pathologist is that the pathologist would never relinquish its moral and professional responsibilities for a botched diagnosis, nor the hospital its clinical and reputational ones. Nevertheless, the patient, the pathologist and the hospital can avoid financial stress with a malpractice insurance policy, whose cost reflects the guarantees and liability clauses the supplier is contractually bound to assume, the quality of the clinical practices in the hospital and the professional standing of the pathologist.

Autonomy as moral agency. Moral agency is understood as the capability of an individual to choose behaviour —make "moral choices" based on some notion of right and wrong— and being accountable for the harmful consequences of those choices. Depending on the requirements one imposes on the decision-making process one can debate whether or not, or in what sense animals, children, disabled persons and autonomous artificial systems have moral agency.

The attribution of moral agency to artefacts has been discussed at large. Some opinions are negative: only humans can hold moral agency, not computers, qualified as 'entities', not as 'agents' [8,25,42,29,7]. Others are more favourable [48,44]. For instance, Floridi and Saunders [18] contend that moral agency is a matter of "level of abstraction". They claim that the concept of moral agent does not require the preconditions of exhibiting free will, mental states or responsibility. Their guidelines for "agenthood" are: *interactivity* (response to stimulus by change of state), *autonomy* (ability to change state without stimulus) and *adaptability* (ability to change the transition rules by which state is changed) at a given level of abstraction. Thus, under these assumptions, moral agency may hold for AI systems as well.

From a classical Multi-Agent Systems perspective, Falcone and Castelfranchi [16] contend that it is possible to analyse the adjustable autonomy of an agent both by considering the level of delegation allowed to the contractor (agent) by the principal, and the possibility for the contractor itself to adjust its own autonomy by restricting or by expanding the received delegation. Falcone and Castelfranchi claim that "in studying how to adjust the level of autonomy and how to arrive to a dynamic level of control, it could be useful an explicit theory of delegation able to specify different kinds and levels of autonomy".

Essentially, what matters in this form of delegation of autonomy is not only the debatable possibility of the "personhood" or even "legal personhood" ([27,1]). The issue that truly matters is to determine whether an artificial system would have, *de facto*, the responsibility for any harm its actions may produce. In other words, in order to mange risk, the funda-

mental problem in artificial moral agency is the proper assessment and allocation of liability in order to render such agent accountable.

Dealing with harm: liability and reparation. We claimed that autonomy is the source of fundamental AI risk and we mentioned two forms of autonomy in AIS: delegation and moral agency. In both cases there are essentially two risks: misuse and malfunction of the artefact.
On the surface, both forms of autonomy can be addressed in a similar way because the ex-post risk-management process starts in both cases by identifying *liability*. That is, assessing the level of harm and attributing responsibility (recall Fig. 2). Once liability (harm and responsibility) is established, a process of reparation can be activated, which is essentially two-fold: compensation of damage and mitigation.
While this ex-post claim management process applies in most situations, its actual execution needs to take into account three salient considerations. The first one is that the impact of the misuse or the malfunction —that is, the severity of each casualty (or claim), the number of casualties and the expected reparation costs— needs to be articulated in the materialisation of each hazard (a given "claim") both to assess actual and potential liabilities, and to devise commensurable reparation processes. The second one is that the conventions and principles to assess liability and address reparation may differ substantially depending on the domain of use of the AIS (health, transportation, civil rights, defence). Finally, in order to implement the claim settlement process (assessment, attribution and reparation), one needs to account for the socio-legal environment where harm is taking place.
Fortunately, it is often the case that in order to establish liability (harm and responsibility) and render guilty parties accountable (blame assignment, reparation measures and enforcement), one can rely on conventional means like contracts, regulations, oversight and insurance, following a conventional due process. However, for some special cases of AI misuse or malfunction, new *ad-hoc* due processes might be needed. Such new due processes ought to be responsive to the impact and domain considerations mentioned above. Moreover, some new non-standard cases will need a more sophisticated analysis of the"value chain" of stakeholders involved in order to properly establish collective and individual liability, as well as the respective reparation processes.[15]

8 Risk control and value engineering.

Although in this paper we circumvent further elaboration of the risk management process for autonomous AIS, we turn our attention, however briefly, to a time-honoured tool to address autonomy, which is *governance*. Instead of reacting to claims, governance assumes a proactive role. Governance prevents harm by imposing restrictions or dissuading undesirable behaviour, while also creating incentives and facilitating courses

[15] See *Recital 20* of the Artificial Intelligence Act [15]) that acknowledges the need of refining such notions and links the overall process to compliance and a better understanding of the "AI value chain".

of action to foster desirable behaviour. This proactive role can be explicitly linked with ethical considerations (as we will argue below).
Conventional governance principles and practices are relevant, once again, to the management of fundamental AI risk. Either by adapting conventional governance means —like standards, best practices, rules and regulations as well as enforcement and oversight frameworks— or, by developing analogous new ones [9]. Nevertheless, one can also look into the unconventional idea of making use of AI to govern artificial autonomous systems. One particular approach to the governance of AI autonomy is to rely on ethics as a form of control. The gist of this approach is sketched in Fig.7.

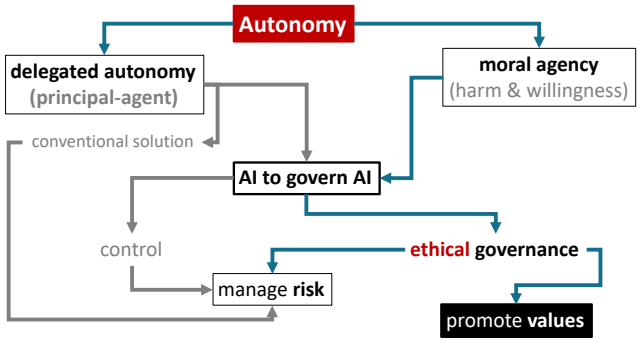

Fig. 7. While conventional practices and means can be adapted judiciously to the governance of artificial delegated autonomy, the deployment of AI means to govern moral autonomy in AIS is a distinct possibility. One venue in this approach is to design AIS that are aligned with human values.

Although AI based AIS governance in general, and ethics-based AI governance in particular, can be approached from several fronts, there are two distinct but complementary salient subproblems: (i) imbuing ethical notions into the self-governance of autonomous agents and (ii) imbuing ethical notions into the governance of the collective interactions in online hybrid human AI social systems [36,37,30].
The general problem of ethics-based AI governance can be framed as a *value alignment problem*: to find ways of guiding behaviour (individual or collective) towards outcomes that are consistent with some values. Value alignment, in the context of AI, can be made explicit as the problem of"designing autonomous artificial systems whose behaviour is objectively aligned with explicit human values" ([46,47,45,19,53]); VAP, for brevity.
VAP can be approached as an engineering problem: how to implement the means to govern the behaviour of an AIS so that the satisfaction of an explicit set of values can be objectively assessed [41]. By acknowledging the correspondence of harm and value, VAP translates, for the purpose of risk-management, into the challenge of designing and imple-

menting governance devices into AIS so that explicit risks are objectively avoided.[16]
What we are in fact claiming is that if one can align an AIS to avoid an explicit risk, it is because the degree to which that risk is avoided in that AIS can be objectively assessed; hence the actual harm can be objectively assessed. That is, risk-bound AIS can be designed if value-aligned AIS can.[17]
This interpretation of VAT from a risk-management perspective does not solve the general problem of fundamental AI risk but it is a promising incursion in that direction: It provides the basis for the design of risk-bound AIS those for which liability can be objectively assessed, accountability can be objectively attributed and the corresponding reparation procedures can be put in place.
This last observation leads to another argument in favour of addressing risk-management as values-based governance: the possibility not only of addressing liability, by avoiding objective harm, but actually achieving benefits by objectively accruing value [34].

9 Closing remarks

We understand that the commitment of the AI community should be to responsible AI development. We propose to instrument this responsibility through a discerning attention to the management of AI risks.

1. We have approached AI risks from a risk-management perspective and identified three main types of risk that are associated with AI (*inertial, disruptive* and *fundamental*). We sketched a tentative strategy to address each of them (vigilant, reactive and preventive).
2. In particular, we are convinced that the AI community has the salient responsibility of addressing the fundamental risks associated with autonomy in AI artefacts. We argue that the challenge of developing *risk-bound AIS* (as a value engineering problem) is a sensible step in that direction.
3. We advocate for a long-term perspective: The judicious strategy to deal with AI risk —including latent risk however large it may be— is to develop institutional frameworks to design and implement AI risk control policies. Risk control policies that should look into ways of (i) developing risk-commensurable governance that is responsive to differentiated risks, and (ii) enabling reliable national and international due processes to deal with AI induced harm, including means to render risk-enhancing stakeholders accountable.
4. The fact that conventional risk-management practices may apply to AI risks provides grounds for cautious confidence, suggests guides

[16] In not all that dissimilar spirit, there have been three other proposals to govern autonomy in which values play some role: superalignment, constitutional AI and risk-level safeguards for "agentic AI" ([12,2,20,21,33,6,5].

[17] Of course this claim can only be validated up to the properties of the engineering process or the methodology that is used to engineer that particular risk [37,34,35].

for their prudent adaptation to AI, and motivates a road-map for the type of research, institutional and market developments needed to avert extreme risks.
5. AI development and use comes with unavoidable risks, but also with benefits. Whatever strategy to contend with AI risk we choose, it should encompass a sound strategy to offset any potential harm with actual benefits.
   In this respect, we urge our colleagues to invest a systematic effort in the development of AI artefacts that align with human values. We argue that value aligned AIS, in addition to their risk-taming role, can be designed to achieve objective social benefits.

In sum, the design of value-aligned artificial intelligence systems is at the foundation for the development of beneficial AI.

Acknowledgments

We wish to thank Jean-Gabriel Ganascia, Ramon López de Mántaras, Mustafa Hashmi, Ho-Pun Lam, and Louis de Koker for their helpful comments. Research for his paper is supported by EU (Horizon-EIC-2021-Pathfinder challenges-01) Project VALAWAI 101070930; the EU Next Generation EU/PRTR program; the Spanish (MCIN/AEI) project VAE TED2021-131295B-C31, CSIC's (Bilateral Collaboration Initiative i-LINK-TEC) project DESAFIA2030 BILTC22005 and the IDT Research Group of Excellence UAB-CSIC (SGR 00532 2022-24).

References

1. Anderson, M., Anderson, S.L.: Machine ethics: Creating an ethical intelligent agent. AI Magazine **28**(4), 15–15 (2007)
2. Bai, Y., Kadavath, S., Kundu, S., Askell, A., Kernion, J., Jones, A., Chen, A., Goldie, A., Mirhoseini, A., McKinnon, C., et al.: Constitutional ai: Harmlessness from ai feedback (2022). https://doi.org/10.48550/arXiv.2211.2212.08073
3. Bales, A., D'Alessandro, W., Kirk-Giannini, C.D.: Artificial intelligence: Arguments for catastrophic risk. Philosophy Compass **19**(2), e12964 (2024)
4. Bengio, Y., Hinton, G., Yao, A., Song, D., Abbeel, P., Darrell, T., Harari, Y.N., Zhang, Y.Q., Xue, L., Shalev-Shwartz, S., et al.: Managing extreme AI risks amid rapid progress. Science **384**(6698), 842–845 (2024)
5. Bowman, S.R., Hyun, J., Perez, E., Chen, E., Pettit, C., Heiner, S., Lukošiūtė, K., Askell, A., Jones, A., et al.: Measuring progress on scalable oversight for large language models (2022). https://doi.org/10.48550/arXiv.2211.03540
6. Brenneis, A.: Assessing dual use risks in ai research: necessity, challenges and mitigation strategies. Research Ethics **0**(0), 17470161241267782 (0). https://doi.org/10.1177/17470161241267782

7. Brożek, B., Janik, B.: Can artificial intelligences be moral agents? New ideas in psychology **54**, 101–106 (2019)
8. Bryson, J.J., Diamantis, M.E., Grant, T.D.: Of, for, and by the people: the legal lacuna of synthetic persons. Artificial Intelligence and Law **25**(3), 273–291 (Sep 2017). https://doi.org/10.1007/s10506-017-9214-9, https://doi.org/10.1007/s10506-017-9214-9
9. Casanovas, P., Hashmi, M., de Koker, L., Lam, H.P.: A three steps methodological approach to legal governance validation (2024). https://doi.org/10.48550/arXiv.2407.20691
10. Castelfranchi, C., Falcone, R.: From automaticity to autonomy: the frontier of artificial agents. Agent autonomy pp. 103–136 (2003)
11. Chalmers, D.J.: The singularity: A philosophical analysis. Science fiction and philosophy: From time travel to superintelligence pp. 171–224 (2016)
12. Christian, B.: Alignemt Problem: machine learning and human values. W W Norton, S.l. (2021), oCLC: 1233266753
13. Daughety, A., Reinganum, J.: Economic analysis of products liability: Theory, pp. 69–96. Edward Elgar Publishing (11 2013). https://doi.org/10.4337/9781781006177.00011
14. Daughety, A.F., Reinganum, J.F.: Markets, torts, and social inefficiency. The RAND Journal of Economics **37**(2), 300–323 (2006). https://doi.org/https://doi.org/10.1111/j.1756-2171.2006.tb00017.x
15. European Commission: Regulation (eu) 2024/1689 of the european parliament and of the council of 13 june 2024 laying down harmonised rules on artificial intelligence (artificial intelligence act) (text with eea relevance) pe/24/2024/rev/1, https://eur-lex.europa.eu/eli/reg/2024/1689/oj
16. Falcone, R., Castelfranchi, C.: Levels of delegation and levels of adoption as the basis for adjustable autonomy. In: Congress of the Italian Association for Artificial Intelligence. pp. 273–284. Springer (1999)
17. Feigenbaum, E.A.: Stories of AAAI–Before the Beginning and After: A Love Letter. AI Magazine **26**(4), 30–30 (2005)
18. Floridi, L., Sanders, J.W.: On the morality of artificial agents. Minds and Machines **14**(3), 349–379 (2004)
19. Gabriel, I.: Artificial intelligence, values, and alignment. Minds and Machines **30**(3), 411–437 (2020). https://doi.org/10.1007/s11023-020-09539-2, https://doi.org/10.1007/s11023-020-09539-2
20. Ganguli, D., Askell, A., Schiefer, N., Liao, T.I., Lukošiūtė, K., Chen, A., Goldie, A., Mirhoseini, A., Olsson, C., et al.: The capacity for moral self-correction in large language models (2023). https://doi.org/10.48550/arXiv.2302.07459
21. Ganguli, D., Lovitt, L., Kernion, J., Askell, A., Bai, Y., Kadavath, S., Mann, B., Perez, E., Schiefer, N., Ndousse, K., et al.: Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned (2022). https://doi.org/10.48550/arXiv.2209.07858

22. Giudici, P., Centurelli, M., Turchetta, S.: Artificial intelligence risk measurement. Expert Systems with Applications **235**, 121220 (2024). https://doi.org/https://doi.org/10.1016/j.eswa.2023.121220, https://www.sciencedirect.com/science/article/pii/S0957417423017220
23. Hawking, S., Russell, S., Tegmark, M., Wilczek, F.: Stephen hawking: 'transcendence looks at the implications of artificial intelligence - but are we taking ai seriously enough?'. The Independent. May 4, 2014. (Retr. Oct 2024)
24. Hendler, J.: Avoiding Another AI Winter. Intelligent Systems, IEEE **23**, 2–4 (04 2008). https://doi.org/10.1109/MIS.2008.20
25. Himma, K.E.: Artificial agency, consciousness, and the criteria for moral agency: what properties must an artificial agent have to be a moral agent? Ethics and Information Technology **11**(1), 19–29 (2009). https://doi.org/10.1007/s10676-008-9167-5, https://doi.org/10.1007/s10676-008-9167-5
26. Hua, X., Spier, K.E.: Holding platforms liable. HKUST Business School Research Paper No. 2021-048, (June 3, 2022) (2022). https://doi.org/10.2139/ssrn.3985066, https://dx.doi.org/10.2139/ssrn.3985066
27. IEEE: Ethically aligned design, version 2 (2017), https://ethicsinaction.ieee.org/, [Online] Retrieved 13 June 2019
28. International Standards Office (ISO) and International Electrotechnical Commission (IEC): International standard iso/iec 23894: Information technology — artificial intelligence — guidance on risk management (first edition) (2023), https://cdn.standards.iteh.ai/samples/77304/cb803ee4e9624430a5db177459158b24/ISO-IEC-23894-2023.pdf
29. Johnson, D.G.: Computer systems: Moral entities but not moral agents. Ethics and information technology **8**, 195–204 (2006)
30. King, T.C., De Vos, M., Dignum, V., Jonker, C.M., Li, T., Padget, J., van Riemsdijk, M.B.: Automated multi-level governance compliance checking. Autonomous Agents and Multi-Agent Systems pp. 1–61 (2017)
31. Lenat, D.B., Feigenbaum, E.A.: On the Thresholds of Knowledge. In: IJCAI. vol. 87, pp. 1172–1176 (1987)
32. Lockey, S., Gillespie, N.M., Holm, D., Someh, I.A.: A review of trust in artificial intelligence: Challenges, vulnerabilities and future directions. In: 54th Hawaii International Conference on System Sciences, HICSS 2021, Kauai, Hawaii, USA, January 5, 2021. pp. 1–10. ScholarSpace, (Retr. Oct 2024) (2021), https://hdl.handle.net/10125/71284
33. METR (Model Evaluation and Threat Research): Responsible scaling policies (rsps), https://metr.org/blog/2023-09-26-rsp/, (retrieved 18 Oct 2024)
34. Noriega, P., Plaza, E.: 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, pp. 130–139. CEUR-WS.org, (Ret. Oct 2024) (2022), https://ceur-ws.org/Vol-3531/SPaper_10.pdf
35. Noriega, P., Plaza, E.: On Autonomy, Governance, and Values: An AGV Approach to Value Engineering. In: Osman, N., Steels, L. (eds.) Value Engineering in Artificial Intelligence. pp. 165–179. Springer Nature Switzerland, Cham (2024). https://doi.org/https://link.springer.com/chapter/10.1007/978-3-031-58202-8_10
36. Noriega, P., Verhagen, H., Padget, J., d'Inverno, M.: Design Heuristics for Ethical Online Institutions. In: Ajmeri, N., Morris Martin, A., Savarimuthu, B.T.R. (eds.) Coordination, Organizations, Institutions, Norms, and Ethics for Governance of Multi-Agent Systems XV. pp. 213–230. Springer International Publishing, Cham (2022)
37. Noriega, P., Verhagen, H., Padget, J., d'Inverno, M.: 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. pp. 77–94. Springer Nature Switzerland, Cham (2023)
38. OECD: OECD Framework for the Classification of AI systems. OECD Digital Economy Papers (323) (2022). https://doi.org/https://doi.org/https://doi.org/10.1787/cb6d9eca-en, https://www.oecd-ilibrary.org/content/paper/cb6d9eca-en
39. OECD: Explanatory memorandum on the updated OECD definition of an AI system (2024). https://doi.org/https://doi.org/https://doi.org/10.1787/623da898-en, https://www.oecd-ilibrary.org/content/paper/623da898-en, (Retrieved Sep. 2024)
40. Orwat, C., Bareis, J., Folberth, A., Jahnel, J., Wadephul, C.: Normative challenges of risk regulation of artificial intelligence and automated decision-making. CoRR **abs/2211.06203** (2022). https://doi.org/10.48550/ARXIV.2211.06203, https://doi.org/10.48550/arXiv.2211.06203
41. Osman, N., Steels, L. (eds.): Value Engineering in Artificial Intelligence - First International Workshop, VALE 2023, Krakow, Poland, September 30, 2023, Proceedings, Lecture Notes in Computer Science, vol. 14520. Springer, Cham (2024). https://doi.org/10.1007/978-3-031-58202-8
42. Parthemore, J., Whitby, B.: What makes an agent a moral agent? Reflections in machine consciousness and moral agency. International Journal of Machine Consciousness **05**(02), 105–129 (2013). https://doi.org/10.1142/S1793843013500017, https://doi.org/10.1142/S1793843013500017
43. Pascal, B.: Pascal's pensées; introduction by t. s. eliot, translated by w. f. trotter in 1910 (1958)
44. Powers, T.M.: On the moral agency of computers. Topoi **32**(2), 227–236 (2013). https://doi.org/10.1007/s11245-012-9149-4
45. Russell, S.: Of Myths and Moonshine. A conversation with Jaron Lanier, 14-11-14. The Edge (November 2014), https://www.edge.org/conversation/the-myth-of-ai#26015, [Online] Retrieved Oct 2024

46. Russell, S.: Provably beneficial artificial intelligence. The Next Step: Exponential Life, BBVA-Open Mind (2017)
47. Russell, S.: Human compatible: AI and the problem of control. Penguin, UK (2019)
48. Sullins, J.P.: When is a robot a moral agent. International Review of Information Ethics **6**(12), 23–30 (2006)
49. Tabassi, E.: Artificial intelligence risk management framework (ai rmf 1.0) (2023-01-26 05:01:00 2023). https://doi.org/https://doi.org/10.6028/NIST.AI.100-1
50. Thom, R.: Les singularités des applications différentiables. In: Annales de l'institut Fourier. vol. 6, pp. 43–87 (1956)
51. Turing, A.M.: Intelligent machinery, a heretical theory. Philosophia Mathematica **4**(3), 256–260 (1996)
52. Ulam, S.: Tribute to john von neumann. Bulletin of the American mathematical society **64**(3), 1–49 (1958)
53. Vamplew, P., Dazeley, R., Foale, C., Firmin, S., Mummery, J.: Human-aligned artificial intelligence is a multiobjective problem. Ethics and Information Technology **20**(1), 27–40 (2018). https://doi.org/10.1007/s10676-017-9440-6, https://doi.org/10.1007/s10676-017-9440-6