European Journal of Futures Research

RESEARCH ARTICLE Open Access

Is artificial intelligence a hazardous technology? Economic trade-off model

Bodo Herzog[1,2*]

Abstract

Artificial intelligence (AI) demonstrates various opportunities and risks. Our study explores the trade-off of AI technology, including existential risks. We develop a theory and a Bayesian simulation model in order to explore what is at stake. The study reveals four tangible outcomes: (i) regulating existential risks has a boundary solution of either prohibiting the technology or allowing a laissez-faire regulation. (ii) the degree of 'normal' risks follows a trade-off and is dependent on AI-intensity. (iii) we estimate the probability of 'normal' risks to be between 0.002% to 0.006% over a century. (iv) regulating AI requires a balanced and international approach due to the dynamic risks and its global nature.

Keywords Artificial intelligence, Existential risks, Trade-off, Regulation, Efficacy

Introduction

Recent progress in artificial intelligence (AI) has lunched a debate about a technological singularity [26]. The publication of AlphaGo Zero in 2017 and subsequently the release of large language models reveal potential benefits as well as fears [35, 37]. Scientists and notable people in power have signed the following statement [11]: "Mitigating the risk of extinction from AI should be a global priority alongside other societal-scale risks such as pandemics and nuclear war".

Governments are discussing the need of a regulatory approach in order to avoid an AI–dystopia (London ai conference, 2023). The European Union has even approved an Artificial Intelligence Act on 2 February 2024 [15]. Yet, how likely is a collapse of civilization or extinction of humanity due to AI in a century? What is the regulatory trade-off? What are the magnitude and determinants of AI risks?

*Correspondence:
Bodo Herzog
Bodo.Herzog@reutlingen-university.de
[1] ESB Business School, Reutlingen University, Alteburgstr. 150, D-72762 Reutlingen, Germany
[2] RRI Reutlingen Research Institute, Reutlingen University, Reutlingen, Germany

In general, the overall debate encompasses two opposing views. One group argues that a technological singularity exists, presenting mainly risks rather than opportunities [9, 12]. Hence, there is need for an AI embargo. Others suggest a moderate regulatory approach because the benefits outweigh the risks [26]. From the literature we know that catastrophic events "are not amenable to experimental verification — at least not more than once" ([32], p. 298). Existential risks and human extinction are not only difficult to quantify or falsify, but they are also intertwined with complex dynamics across technological, economic, social, and environmental systems. Any existential risk (x-risk) depends on the 'anthropic shadow' [13]. Naturally, modelling catastrophic events require heavy-tail distributions accounting for a substantial degree of uncertainty [49].

In this paper, we study the regulatory trade-off in an interdisciplinary framework, modelling the risks and benefits. Our approach develops a scientific perspective to manage future risks of AI applications. We argue for a balanced awareness and a humbled policy approach. We construct a theory and explore it in a simulation model by applying the Value of a Statistical Life (VSL) theory in order to assess the economic costs associated with AI-induced risks, focusing on non-existential risks and

human extinction. Utilizing a Bayesian methodology, we project total factor productivity (TFP) and refine estimates of the probability of catastrophic events and their corresponding economic impact.

Our findings indicate that x-risks, defined as high risk applications, do not follow a trade-off rather a boundary solution. Prohibition of those applications is reasonable given the difficulty of quantification. Yet, we equally find that the present risks of AI technology are manageable, depending on various scenarios. Similarly to Nordhaus [26] we do not find a singularity, defined as a rapid growth in AI, leading to an ever-increasing and infinite pace of improvement cascades and subsequently risks. The future probability of AI-induced 'normal' risks range from 0.002% to 0.006%. The potential global costs could range from $1.44 trillion to $4.32 trillion by the end of the century. These results highlight the need for a balanced approach, emphasizing the prohibition of high-risk AI applications, while acknowledging the potential benefits of middle to low risk applications. Our research spurs the regulatory debate to a scientific-based direction about the merits and costs of future AI dynamics.

The paper is organized as follows: In the next section we present a literature review. In the subsequent section, we build an analytical model and combine it with a Bayesian time series approach in order to simulate risk probabilities. The discussion section embeds the interpretation of the model results. Finally, we provide concluding remarks.

Literature review

Modelling technological risks is an interdisciplinary endeavor. Although there is a lack of an exact definition, we use a broad terminology of the term existential risks, including human extinction or a major catastrophe [45]. Note, existential risks cannot be studied in a vacuum. Catastrophes evolve in combination to other domains or in chain reactions [2, 19, 43]. The methodological challenges in quantifying the likelihood of existential hazards are studied by Beard et al. [5] and Garibay et al. [16].

The philosophical arguments of a technological dystopia is as old as philosophy. Already Plato (370 BC) claimed that the development of writing degenerates human thinking in future. Contemporary philosophers argue that the probability of humans extinction is imminent due to the probability that we are living in a computer simulation of the world [7, 8]. Those testimonies seem rather science fiction than science, yet even a scientifically well founded Bayesian approach requires a prior probability that is updated with evidence in order to obtain the posterior.

In the context of x-risks, there is sparse evidence about technological related events and the prior probability are disproportionately relevant for a prediction. Although Rees [29] claims x-risks are growing due to quick technological developments, there exist no empirical evidence about human extinction due to technological progress over the past 250 years of industrial revolution; rather the opposite.[1]

A group of researches argue that humans overstate x-risks of new technological developments [44]. In economics there is a historical record of those examples. Distinguished scientists such as Thomas Robert Malthus, John Maynard Keynes, Marvin Minsky or Wassily Leontief have extremely overestimated the risks of future technological developments. Overconfidence and overestimation are inherent notions in psychology and behavioral economics [6, 10, 14, 21].

A few studies are estimating the probability of existential risks [3, 18, 36]. Those studies apply the frequency statistics methodology, which may underestimate the level of sparse events and uncertainty, such as a rare catastrophe. Most studies about existential risks are in the field of global pandemic's. Heuristically, we see a global pandemic each century, such as the Spanish flu in 1918 and the SARS-CoV2 pandemic in 2020. At the same time, we have not seen similar catastrophes caused by technology since the industrial revolution in 1750. Woo [50] argues that we need to incorporate a counterfactual analysis of near-miss events in order to obtain better predictions about rare catastrophes [28, 51].

There are studies addressing the probability of existential risks of AI [4, 17, 23, 25]. Those papers mainly focus on the aggregation of expert opinions. An objection of those estimates are behavioral biases. The study by Grace et al. [17] applies a more rigorous scientific approach. Yet, all studies use a kind of median estimate, which might be inadequate for existential risk modelling according to Turchin [46]. Indeed, estimating x-risks of a future 'artificial general intelligence' (AGI) which does not exist in 2024, should concentrate on tail-risks — that is a low probability event with utmost consequences [22]. Therefore, one can argue that x-risks should be defined by a minimum acceptable level rather than a median risk exposure.

Subjective estimates based on weak priors suffer from biases, such as overconfidence [6, 14, 21]. This insight is not new. Already Yudkowsky [53] discussed the relevance of cognitive biases, such as hindsight bias, heuristic bias, conjunction fallacy, overconfidence and omitted variable bias in judgements of global catastrophic risk assessments. In addition, there is a large difference when

[1] We observe a growing population and scarcity of human labor in the 21st century.

asking for the probability of superhuman intelligence being developed within a century or when asking for the probability of human extinction within a century. The responses from those questions can differ substantially from group to group [23, 34, 36].

There is evidence that predicting the short-term merits or challenges of future technology is more successful than the long-term future developments [41, 42]. Furthermore, the aggregation of expert opinion can diverge from consensus to extreme views. There is evidence that disciplinary views or certain political clusters concentrate its estimates around opposing positions, particularly in a polarized society [31, 38, 39].

Our model framework is close to Nordhaus [26], yet sets itself apart in the following ways: Firstly, we define a novel and tractable optimal control problem. Secondly, we analyze the trade-off in theory. Thirdly, we explore a simulation model with a Bayesian methodology. Lastly, we uncover insights regarding the regulatory design of existential and non-existential risks.

A model

AI technology either augment or replace human tasks [1]. By definition this is leading to an acceleration of economic growth, g. The advancement towards an AGI, however, might pose x-risks to humanity in the long-run [49]. Hence, we have to study the regulatory trade-off of AI's benefits and costs to humanity.

Let $S(t)$ denote the production function of economic stability $S(t) = s_0 e^{g(L(t))t}$, where AI technology is economically subsumed to augment innovation and welfare over time. The function $L(t) = l_0 e^{\xi t}$ represents the exponential rise in AI intensity over time, t. The parameter ξ denotes the growth rate of future AI technology. Furthermore, we assume that the economic growth rate is linearly dependent on AI intensity: $g(L(t)) = \bar{g}L(t)$. We define the relationship between both functions in form of an ordinary first-order differential equation:

$$\frac{dS(t)}{dt} = -L(t). \tag{1}$$

Intuitively, an intensifying usage of AI technology, $L(t)$, replaces tasks and decrease the growth rate of economic stability $dS(t)/dt$ because it exposes the economy to future catastrophes.

We define a function $U(L)$ that denotes a relationship between AI intensity and economic utility. This function has the derivatives of $dU(L)/dL > 0$ and $d^2U(L)/dL^2 < 0$. Moreover, AI applications create disutility in regard to growing risks, denoted by the function $P(L)$.[2]

We describe the risks by defining a probability of survival $P^{-1}(L(t)) = \delta e^{-\delta L(t)}$ considering the accumulation of risks with a right-skewed density function [47]. The social welfare of the society depends on both the potentials of AI, $U(L(t))$, and the risks posed by AI, $P(L(t))$. The welfare function is defined by

$$W = W(U(L(t)), P(L(t))), \tag{2}$$

where we assume $dW/dU > 0$, $dW/dP < 0$, $d^2W/dU^2 < 0$, $d^2W/dP^2 < 0$ and $d^2W/dUdP = 0$. This specification postulates that marginal utility of stability has a positive but diminishing contribution to welfare. In contrast, the marginal utility of risks are negative over time.

Since both functions depend on $L(t)$, we treat this parameter as the control variable in our optimization problem. The degree of stability, $S(t)$, in Eq. (1) is the state variable. Note, in our simulation model, we focus on the probability in the end. We do not focus on the order of magnitude of a catastrophe [13]. In our simulation approach, we include the magnitude by utilizing the concept of value of a statistical life (VSL).

Optimal control problem

Suppose a public authority is regulating the future of AI. Let us compute the optimal degree of AI intensity, $L(t)$, over a time horizon $[0, T]$. The optimal control problem is

$$\begin{aligned} \max_{L(t)} \quad & \int_0^T W(U(L(t)), P(L(t)))dt \\ \text{s.t.} \quad & \frac{dS(t)}{dt} = -L(t) \\ \text{and} \quad & S(0) = s_0 \qquad S(T) \geq 0 \qquad (s_0, T \quad \text{given}) \end{aligned} \tag{3}$$

The so-called Ramsey approach avoids a discount factor in the integrand. The regulatory authority has discretion of selecting a certain level of stability, $S(T)$, subject to a reasonable restriction that it is nonnegative. The Hamiltonian of this problem is

$$H = W(U(L(t)), P(L(t))) - \lambda(t)L(t). \tag{4}$$

Maximizing H with respect to the control variable $L(t)$ by setting its first derivative equal to zero yields

$$\frac{\partial H}{\partial L(t)} = W_U \frac{dU(L)}{dL} + W_P \frac{dP(L)}{dL} - \lambda(t) = 0 \tag{5}$$

where $W_U = dW/dU$ and $W_P = dW/dP$. To make sure that we obtain a maximum, we check the second derivative (Appendix A). To elicit more information about $L(t)$ from Eq. (5), we look into the time path of λ. The maximum principle tells us the law of motion for λ. We obtain

[2] $dP(L)/dL > 0$ and $d^2P(L)/dL^2 > 0$.

$$\frac{d\lambda}{dt} = -\frac{\partial H}{\partial S(t)} = 0. \tag{6}$$

This condition implies $\lambda(t) = c$. Defining the constant c, we use the transversality condition. With a truncated terminal line the condition takes the form $\lambda(T) \geq 0$ and $S(T) \geq 0$ as well as $\lambda(T)S(T) = 0$. It is evident that $\lambda(T) = 0$. Additionally, due to Eq. (6) for $\lambda(t)$, we obtain that it is zero for all t. With $\lambda(t) = 0$, Eq. (5) reduces to

$$\frac{\partial H}{\partial L(t)} = W_U \frac{dU(L)}{dL} + W_P \frac{dP(L)}{dL} = 0 \tag{7}$$

which can be solved for an optimal path in L^*. If the equation is independent of time t, the solution is constant over time: $L^*(t) = L^*$, where L^* denotes the optimal acceptable degree of AI intensity. Whether this solution is acceptable from the standpoint of the benefits $S(T) \geq 0$ remains a matter to be settled. Prior to continuing, it is useful to examine the economic meaning of Eq. (7) in general.

The first term, $W_U(dU(L)/dL)$, measures the marginal effect of a change in AI intensity on the societal welfare. It represents the marginal utility of AI usage through its contribution to economic stability. The second term, $W_P(dP(L)/dL)$, expresses the marginal disutility derived from AI risks. Therefore, Eq. (7) determines the optimal trade-off of a regulatory problem. Specifically, this equation represents the trade-off between the merits and costs of AI regulation.

Theorem 1 *The regulatory trade-off is determined by AI intensity $L(t)$ in Eq. (7).*

Proof Solution of the optimal control problem (3) yields Eq. (7). □

It remains to investigate whether L^* satisfies the restriction $S(T) \geq 0$. Next, we explore the state path of stability $S(t)$. Integrating the first-order differential equation, we obtain

$$S(t) = -t \cdot L + k \tag{8}$$

where k is an arbitrary integration constant. For $t = 0$, we assume $S(t) = S_0$. Hence, the optimal state path is of $S(t) = S_0 - t \cdot L$. The optimal degree of stability $S^*(t)$ at any time hinges on the magnitude of AI's intensity L. Of course, the trade-off has to balance not merely AI's costs and benefits, it equally has to consider the costs of regulation in the end.

Further extension

The model above is assuming a certain AI intensity, which is a flow variable over time. It does not accumulate but (x-) risks do [49]. What happens if $L(t)$ does accumulate?

Suppose the costs of AI usage is lasting and follows the change of the disutility function

$$\frac{dP(t)}{dt} = \alpha L(t) - \beta A(t) - \gamma P(t), \tag{9}$$

where $\alpha, \beta > 0$ and $0 < \gamma < 1$. The first term models the growth in disutility due to higher AI intensity. The second term, $\beta A(t)$, denotes an insurance against x-risks. Indeed, $A(t)$ represents the willingness-to-protect against growing x-risks. The last term expresses the observation, that a higher magnitude of x-risks do curb due to the law of diminishing disutility. Note, we do not assume any explicit functions for $L(t)$ now. In essence, the regulatory architecture reduces the potential damages and costs over time. To complete the model, we have to take the variable $A(t)$ into consideration and rewrite Eq. (1) as

$$\frac{dS(t)}{dt} = -A(t) - L(t). \tag{10}$$

The optimal control problem is transformed to

$$\begin{aligned}
\max_{L(t)} \quad & \int_0^T W(U(L(t)), P(L(t), A(t), P(t)))dt \\
\text{s.t.} \quad & \frac{dP(t)}{dt} = \alpha L(t) - \beta A(t) - \gamma P(t) \\
& \frac{dS(t)}{dt} = -A(t) - L(t) \\
& P(0) = P_0 > 0 \qquad P(T) \geq 0 \\
& S(0) = S_0 \qquad S(T) \geq 0 \qquad (S_0, P_0, T \text{given}) \\
\text{and} \quad & L(t) \geq 0 \qquad 0 \leq A(t) \leq \bar{A}
\end{aligned} \tag{11}$$

Two aspects are of interest. First, the terminal values of the disutility of risks, $P(t)$, and the stability, $S(t)$, are left free in the future at time T. Second, both control variables, the intensity of AI, $L(t)$, and the insurance payment, $A(t)$, are confined in certain ranges. For $L(t)$, the interval is of $[0, \infty)$, which means an AI embargo if $L(t) = 0$ or no policy intervention if $L(t) \to \infty$. For $A(t)$, the control region is of $[0, \bar{A}]$, where $\bar{A}$ denotes the maximum agreement (insurance level) of a global authority.

By writing the Hamiltonian function, we obtain

$$H = W(U(L_t), P(L_t, A_t, P_t)) + \lambda_P[\alpha L_t - \beta A_t - \gamma P_t] - \lambda_S[A_t + L_t] \tag{12}$$

where the subscripts of $P(t)$ and $S(t)$ to each costate variable λ_i indicate the associated state variables. We maximize H with respect of $L(t)$ and using the Kuhn-Tucker

condition of $\partial H/\partial L(t) \leq 0$ together with the complementary-slackness condition $L(t)(\partial H/\partial L) = 0$. We rule out the case of $L(t) = 0$ and logically postulate some $L(t) > 0$.

The complementary slackness condition satisfies

$$\frac{\partial H}{\partial L(t)} = W_U \frac{dU(L)}{dL} + \lambda_P \alpha - \lambda_S = 0 \tag{13}$$

The second derivative is negative and so H is maximized (Appendix B). In addition, we maximize H with respect to $A(t)$:

$$\frac{\partial H}{\partial A} = -\beta\lambda_P - \lambda_S \tag{14}$$

Note, $A(t)$ is restricted to the closed set $[0, \bar{A}]$. If $\partial H/\partial A$ is negative, the left-side of the boundary solution is $A^* = 0$. If $\partial H/\partial A$ is positive, we obtain $A^* = \bar{A}$. In summary,

$$\beta\lambda_P + \lambda_S \begin{pmatrix} > \\ < \end{pmatrix} 0 \qquad \Rightarrow A^* = \begin{pmatrix} 0 \\ \bar{A} \end{pmatrix}. \tag{15}$$

Complementary slackness together with Eq. (13) exhibits

$$\lambda_S = W_U * \frac{dU}{dL} + \alpha\lambda_P. \tag{16}$$

Using the last condition in Eq. (15), we obtain

$$W_U * \frac{dU}{dL} \begin{pmatrix} > \\ < \end{pmatrix} - (\alpha + \beta) * \lambda_P \qquad \Rightarrow A^* = \begin{pmatrix} 0 \\ \bar{A} \end{pmatrix} \tag{17}$$

The optimal degree of insurance critically depends on λ_P.

Lemma 1 *Insurance against x-risks has no interior solution*: $A \notin (0, \bar{A})$.

Proof Consider the dynamic equations of the following costate variables:

$$\dot{\lambda_P} = -\frac{\partial H}{\partial P} = -W_P + \lambda_P\gamma \tag{18}$$

$$\dot{\lambda_S} = -\frac{\partial H}{\partial S} = 0 \qquad \Rightarrow \lambda_S = constant \tag{19}$$

If A^* is an interior solution, then $\beta\lambda_P + \lambda_S = 0$. Since λ_S is a constant, this equation shows that λ_P is a constant too. In turn, we obtain

$$\dot{\lambda_P} = 0 \qquad \Rightarrow \qquad \gamma\lambda_P = W_P \tag{20}$$

This implies W_P to be constant. Since the welfare function W is monotonic in P, there can only be one value of P that would make W_P be a constant. Given $P_0 = P(T) > 0$, the transversality condition reveals

$$P(T)\lambda_P(T) = 0. \tag{21}$$

A positive $P(t)$ implies $\lambda_P(T) = 0$ for all $t \in [0, T]$. Consequently, a zero value for λ_P implies an interior solution by Eq. (16). Given $W_U(dU/dL) = 0$, we have a contradiction to the assumptions of W_U and dU/dL are both positive. Consequently, an interior solution for A^* must be ruled out. □

Theorem 2 *The optimal policy encompasses a boundary solution for insuring x-risks and requires to define the optimal AI intensity L^*.*

This boundary solution is related to $W_U \frac{dU(L)}{dL} = \lambda_S - \alpha\lambda_P$. Intuitively, the effect of regulating AI's intensity pass-through utility, $W_U(dU(L)/dL)$. Both policies need to be equated to the respective shadow prices for AI's stability, measured by λ_S, adjusted by the shadow price of risks $-\alpha\lambda_P$. Both policies differ such that

$$A^* = \begin{pmatrix} 0 \\ \bar{A} \end{pmatrix} \Leftrightarrow \lambda_S \begin{pmatrix} > \\ < \end{pmatrix} - \beta\lambda_P. \tag{22}$$

In the first case it is not worthwhile to expand an AI insurance scheme because AI's impact on stability is less than the risks. On the contrary, implementing an insurance scheme is recommended, if the trade-off outweighs the risks. Distinguishing both cases, we find that the parameter β, which measures the efficacy of regulation plays an essential role.

Lemma 2 *The optimal regulatory degree depends on regulatory efficacy β.*

Proof By using Eq. (22). □

We highlight three insights: (a) regulation of 'normal' risks follow an economic trade-off. (b) insurance of x-risks do not follow a trade-off. Indeed, insurance does not exist for x-risks and only have a boundary solution. Either no insurance $A^* = 0$ but consequently high risks or full insurance $A^* = \bar{A}$, which does not exist, such as an insurance for a nuclear power plant. (c) the efficacy of regulation is essential. For global AI services the regulatory approach requires a global level-playing field.

There are policy implications of our theory. Although, we cannot pin down the magnitude of regulation, our findings suggest that regulation within accumulating damages requires to define the level of AI intensity and

risks a society is willing to tolerate over time. With the knowledge of our model, we conclude that identifying potential high risks applications is a sensible regulatory approach. However, the design of the EU AI-act is not necessarily optimal for low to middle risk applications because in that corridor perhaps benefits outweigh regulatory costs.

Simulation model

The simulation of our model is based on the theory of the value of a statistical life (VSL). This is a standard concept in a cost-benefit analysis to quantify the monetary value associated with reducing the risk of death or extinction. The VSL is not the value of an individual's life per se, but rather the value society places on reducing the risk of death (e.g., one in a million). In other words, the VSL represents how much people are willing to pay to reduce their risk of dying. It is derived from observing people's behaviors and choices in situations involving risk, such as job markets and consumer behavior. The VSL is computed on observable data by how much extra pay workers require to accept higher risks of death or by surveys asking individuals how much they are willing to pay for small reductions in risks.

Today, the VSL is used by policymakers to evaluate the benefits of regulations and interventions that reduce mortality risks. With this concept policy can decide whether new regulations are worth implementing. The OECD [27] suggest VSL ranges between \$1 to \$10 million, depending on low- and high-income countries. For instance, if a policy is projected to save 100 lives and the VSL is \$10 million, the benefit is \$1 billion. If the cost of implementing the policy is less than \$1 billion, it is considered worthwhile.

Note, the VSL estimates vary significantly based on the approach and context. For instance, a study comparing different methodologies found average VSL estimates for the US ranging from \$4.47 million to \$10.63 million depending on the model [40].

We follow the literature and use maximum estimates in order to derive a conservative upper bound. A early study by Viscusi & Aldy [48] estimated the VSL across different studies and countries. They were finding a range from \$1 million to \$10 million, with a central tendency around \$6 million to \$7 million. A comprehensive meta-analysis by the OECD synthesized numerous VSL studies, reflecting a global perspective. This meta-analysis provides valuable insights into how VSL estimates vary by country and risk perceptions [27]. The VSL numbers range between \$1 to \$10 million.

In addition, there is evidence from the U.S. Environmental Protection Agency (EPA), which use a VSL of approximately \$6 to \$9 million (2020 dollars). VSL estimates in European countries tend to be somewhat lower than in the U.S., reflecting differences in income and risk preferences. Typical values range from 1 to 5 million Euro. The European Commission is frequently using values around 2 to 3 million Euro in their regulatory impact assessments. Estimates by Robinson et al. [30] and by the World Health Organization (WHO) range from \$1 million to \$5 million in lower-income countries and up to \$10 million in high-income countries.

We use the formula in Eq. (23) to quantitatively assess the AI induced risks. In order to compute an upper bound, we assume a VSL of \$9 million globally. The benchmark probability in the literature is of 0.001%. The expected cost are computed as

$$\mathbb{E}[C] = prob(risks) \times VSL. \tag{23}$$

For a global population of 8 billion, we obtain for the total VSL \$72 quadrillion (= 8 billion × \$9 million). Consequently, the expected costs are $\mathbb{E}[C] = 0.00001 \times \$72$ quadrillion = \$720 billion. If the global regulatory costs are less than \$720 billion, implementing AI safety measures are recommended according to this benchmark calculation.

The trivial approach above is based on strong assumptions, such as the VSL and the assumption about the probability of catastrophe risk. Next, we extend the benchmark approach by utilizing the major theorem in our analytical theory and combine it with a simulation model. Nonetheless, we stick to the following two assumptions: (a) the VSL is of 9bn US-dollar and (b) the global population is of 8bn.

Our simulation model estimates the probability of risks at the end of this century. The relevant inside of our model is that economic stability is reliant to AI intensity. The simulation utilizes economic data for economic stability. The change in economic stability over time is approximated by the growth rate of total factor productivity (TFP). The TFP data for the United States (US) is downloaded from 1950 to 2034.[3]

In step one of our simulation model, we need to approximate TFP growth by a polynomial function. This polynominal approximation is necessary for a Bayesian time series model in the next step. We specify the approximation by a polynominal function of degree 7 (Fig. 1).

In step two, we build a Bayesian time series model for TFP growth. This model obtains uncertainty bands for TFP growth over the data period of 1950 to 2034. Figure 3 in Appendix C represents the graphical outcome, including relevant upper and lower bounds of TFP growth.

The simulation outcome exhibits potential TFP growth bounds for the US. TFP growth ranges between high

[3] https://fred.stlouisfed.org/

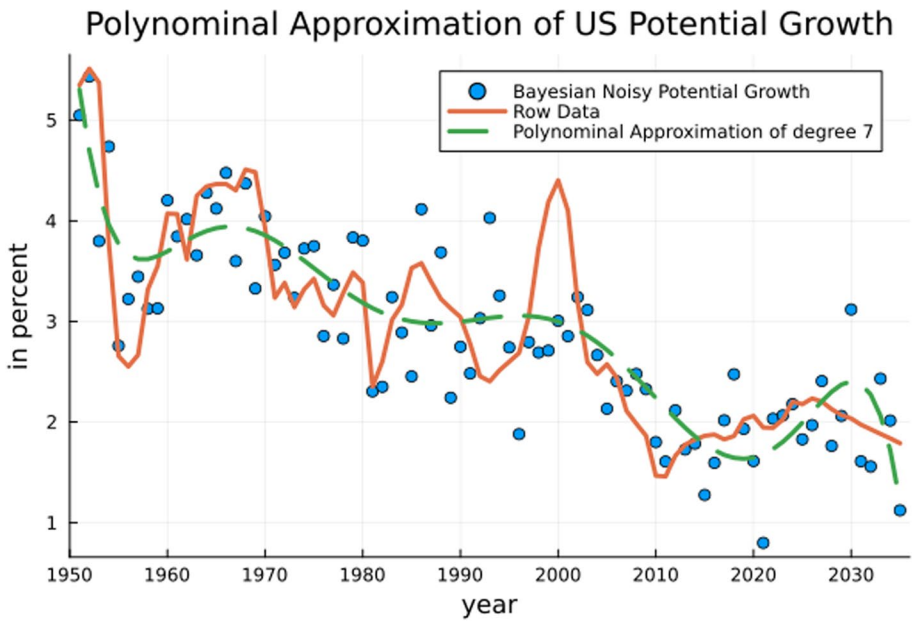

Fig. 1 Simulation data. Source: author

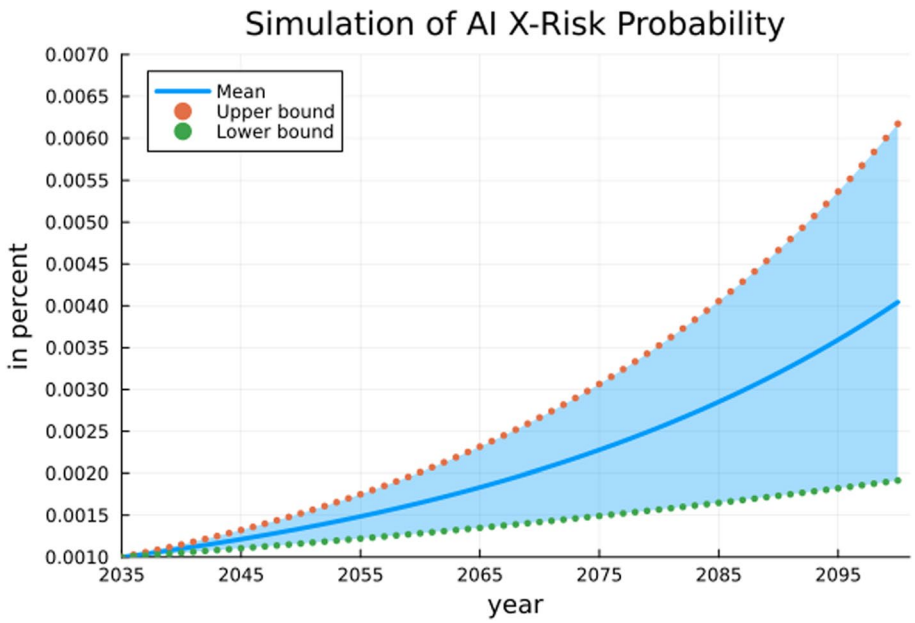

Fig. 2 Simulation of X-risk. Source: author

growth of 2.8%, middle growth of 2.0% and low growth of 1.0%. This range is economically realistic, yet it is derived from our Bayesian simulation model. This model dynamics is applied to forecast the future probabilities of AI related (normal) risks until 2100.

Recall the definition in Eq. (1) that relates the change of economic stability to AI intensity. Moreover, remember that the inverse defines the AI risk probability. Figure 2 simulates the risk probability for three scenarios up to the year of 2100. The numbers range from low risk of 0.002% to middle risk of 0.004% to high risk of 0.006% or levels significantly higher for human extinction (due to heavy tails).

Utilizing this information from our simulation model, we compute the expected total costs or expected damage

Table 1 Accumulation of potential costs of AI-Risks in 2100. Source: author

Simulation Scenario	Trillion US-Dollar
High-Risk Scenario	4.32
Middle-Risk Scenario	2.88
Low-Risk Scenario	1.44

by applying Eq. (23). The results are reported in Table 1. The numbers range from 1.44tr to 4.32tr US-dollar.

The risks and total costs of extreme high-risk applications, such as defined by the EU AI-act, should be prohibited according to our theory because those risks have a boundary solution anyway. Yet, at the same time, the normal scenarios which are based on the numbers above indicate that for applications below the extreme high-risk threshold should refrain from AI regulation. Note, the expected total costs of 4.32tr US-dollar for global damages until 2100 must be balanced to the total future benefits of AI applications until 2100. The current benefits from applications in areas of health, education and industry are significant but do not include the potential benefits over the next 70 years. Most future applications are in its infancy or even not developed. The economic damage of 4.32tr in 2100 in relation to world GDP of 105tr US dollar in 2024 likely outweigh the future potential costs of AI applications over the next 70 years.

Discussion

The nature of the singularity hypothesis proposed by Kurzweil [20] assumes human labor is replaced through artificial intelligence [24, 33]. This requires both task automation and task innovation. Theoretical and empirical evidence by Acemoglu & Restrepo [1] does support task automation. However they find no evidence that humans are obsolete in complex innovation processes. In this domain labor might have a comparative advantage.

Our approach is useful to stimulate future research in that field. Naturally, all of this is speculative, but it helps to better understand why earlier dynamics differ or equal the future. Our interdisciplinary approach has merits and challenges, yet follows a rigorous perspective [5].

The model presents a comprehensive framework for evaluating the trade-off between the benefits and risks associated with increasing intensity of AI technology, particularly in the context of potential x-risks posed by the advancement towards artificial general intelligence. The focus is on balancing the utility derived from AI against the disutility arising from the associated risks.

We observe a nexus between economic stability and AI intensity. The model posits that economic stability, $S(t)$, is intrinsically linked to the intensity of AI, $L(t)$. As AI intensity increases, it accelerates long-term economic growth and enhances economic stability on the one hand. The exponential growth of AI usage over time, however underscores the accelerating pace of technological adoption and (x-)risks on the other hand.

The utility function $U(L)$ and the disutility function $P(L)$ of risks have distinct characteristics. While utility exhibits diminishing marginal returns, risks display increasing marginal disutility. The welfare function incorporates these opposing effects, ensuring that any optimization considers both the positive and negative impacts on societal welfare.

The optimal control problem aims to maximize societal welfare over a given time horizon by determining the optimal AI intensity L^*. The regulation of AI must navigate the delicate balance between fostering technological innovation and safeguarding against potential risks. The model suggests that high AI intensity can drive economic stability, it also necessitates careful regulation to manage the associated risks effectively. The boundary solution for insurance against x-risks indicates that the optimal regulatory policy may involve no insurance $A^* = 0$ or maximum insurance $A^* = \bar{A}$, depending on the efficacy parameter β.

Our findings imply that regulatory policies like the EU AI-Act, which aim to preemptively address extremely high-risk applications, are well founded. However, for low to medium-risk applications, the model suggests that benefits may outweigh the costs, and overly stringent regulations could stifle innovation. A nuanced approach aligns with the broader literature on catastrophe risk management, advocating for preemptive action in high-risk scenarios, while allowing for flexibility in lower-risk spaces.

The main findings of our simulation model support the analytical theory. By utilizing a Bayesian time series model, we provide a structured framework to estimate the future probabilities of catastrophic AI-related events and their associated costs. The concept of VSL is pivotal in our cost-benefit analysis, offering a quantifiable measure to evaluate the worthiness of interventions aimed at reducing mortality risks.

The simulation results are depicted in Fig. 2. We find a variability in potential TFP growth, ranging from high (2.8%) to low growth (1.0%). In addition, we obtain a prediction of risk probabilities from 0.002% in a low-risk scenario to 0.006% in a high-risk scenario, translating to potential costs ranging from $1.44 trillion to $4.32 trillion.

Our findings support the view to prohibit existential-risk applications via stringent regulation. However, at the same time, they indicate for applications below the high-risk threshold, no or low regulation. The projected upper

bound of 'normal' risks of $4.32 trillion by 2100 must be weighed against the anticipated benefits, particularly in sectors such as health, education, and industry. Given the potential benefits and the nascent state of many AI applications today, a balanced regulatory approach is essential to harness AI's positive impact while mitigating its risks.

The estimation of the costs of x-risks are difficult due to high future uncertainties and unknown unknowns. Turchin & Denkenberger [47] suggest to apply a kind of 'Torino scale' of asteroid dangers, ranging from the color white (no risk) to red (extreme risk). Interestingly, they predict AI's x-risks as high (red color) within the next two to three decades, while the x-risks of a pandemic as low (yellow color). This is noteworthy, given we have faced a global pandemic in 1918 and 2020 but we have not seen a technological induced human extinction since the 1st industrial revolution. Thus, the impact assessment of AI technology in the background of technological development of the past relates the perspective.

Similarly, Yudkowsky [52] estimate AI x-risks as high because people conclude too early that they understand the technology. After the public release of ChatGPT almost all interested academics, business persons, or policy-makers claim to understand either the treat or merit of AI technology. In that regard, Yudkowsky [52] was right, "it is very easy for people to think they know far more about Artificial Intelligence than they actually do". Our model intends to estimate the risk without assuming to understand the unknown development of AI technology and without having subjective perceptions about potential risks or benefits. Our estimates are based on sparse parameters and utilizes the value of a statistical life concept.

In addition, there are further objections of the singularity hypothesis. Infinite growth requires energy but energy is limited and is not superabundant. An infinite dynamic would violate basic laws of nature among others the second law of thermodynamics. An open research question is the degree of substitutability between AI and human labor in human-dominated domains, such as innovation, government or regulation of technology.

Future research should focus on refining risk probability estimates and extending the analysis to other countries and regions, considering their specific economic conditions and AI adoption rates. Additionally, exploring alternative models and incorporating further economic data will enhance the robustness of our findings. Understanding the interplay between AI advancements and economic stability remains crucial for developing effective policies that safeguard against catastrophic risks while promoting innovation and growth. Moreover, future research should consider the dynamic accumulation of AI-related risks and the potential long-term impacts on societal welfare. Extending the analytical model to include more granular risk assessments and varying regulatory efficacy across different AI applications could provide deeper insights into optimal regulatory strategies. Additionally, empirical validation of the model's assumptions and predictions would strengthen the robustness and applicability of the findings.

In summary, our model highlights the critical importance of balancing the benefits and risks of AI technology through thoughtful and adaptive regulation. As AI continues to evolve, policymakers ought to remain vigilant in managing its impact on economic stability and welfare, ensuring that the trajectory towards AGI is both beneficial and safe for humanity.

Conclusion

This paper identifies insights and policy conclusions for the design and regulation of artificial intelligence. In order to reduce risks an efficient and dynamic regulatory approach ought to be enforced. The impact of risks and the role of regulation is determined by AI-intensity, regulatory efficacy and substitutability of human labor.

We thoroughly discuss the implications of our combined analytical model and simulation approach. The analytical model explores the economic trade-off between benefits and costs. Our simulation is grounded in the VSL theory and bolstered by Bayesian analysis. This modelling framework offers a comprehensive setup for evaluating the economic implications of AI-induced risks. Projected costs of artificial intelligence underscore the need for targeted regulatory measures, particularly in distinguishing between high- and low-risk applications. Achieving a balance between the potential costs and benefits of AI advancements will be crucial for ensuring a safe and prosperous future. The evolving nature of AI technology and its economic impacts requires continuous research and adaptive policy frameworks to effectively address the shifting landscape of AI risks and opportunities.

Appendix A. Proof of Lemma 1

The second derivative of $\frac{\partial^2 H}{\partial L(t)^2}$ based on the (5) yields:

$sign\frac{\partial^2 H}{\partial L(t)^2} = U_{FF}\left[\frac{dF}{dL}\right]^2 + U_F\frac{d^2F}{dL^2} + U_{PP}\left[\frac{dP}{dL}\right]^2 + U_P * \frac{d^2P}{dL^2} < 0.$

That proofs Lemma 1, due to model assumptions.

Appendix B

The second derivative of Hamiltonian in the extended model yields: $\partial^2 H/\partial L(t)^2 = U_{FF}(dF/dL)^2 + U_F(d^2F/dL^2) < 0.$

Appendix C

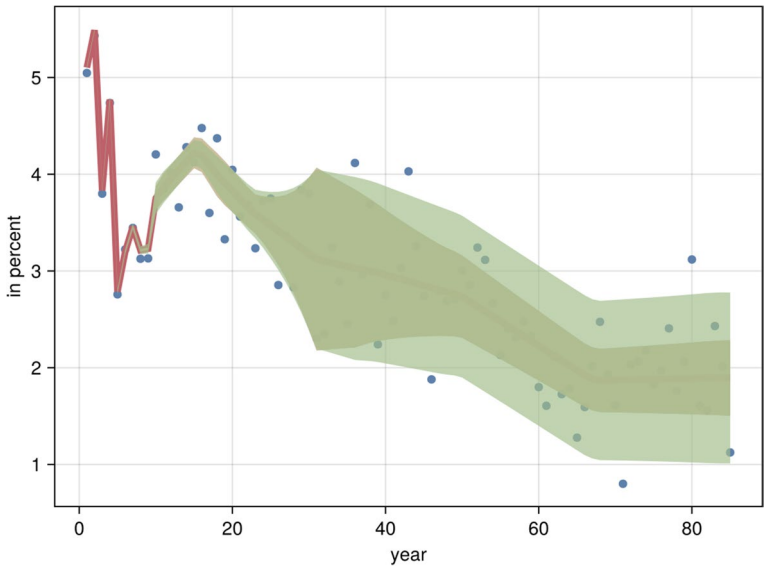

Fig. 3 Simulation of bayesion potential growth until 2100. Note: On the x-axis the zero stands for 1950 and ranges up to 2034. The labelling is different due to simulation. Source: author

Acknowledgements
I would like to express my gratitude for valuable feedback and various comments from conference and seminar participants as well as anonymous reviewers. In addition, I am grateful to the RRI-Reutlingen Research Institute for supporting my research. I hereby declare that I have no conflict of interest, and all remaining errors are in my responsibility.

Authors' contributions
I am the single author.

Funding
Open Access funding enabled and organized by Projekt DEAL.

Availability of data and materials
Yes. Data is publicly available and upon request from the author.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
No.

Received: 25 July 2024 Accepted: 5 August 2024
Published online: 03 September 2024

References

1. Acemoglu D, Restrepo P (2022) Tasks, Automation and the Rise in US Wage Inequality. Econometrica 90(5):1973–2016
2. Avin S, Wintle BC, Weitzdörfer J et al (2018) Classifying global catastrophic risks. Futures 102:20–26. https://doi.org/10.1016/j.futures.2018.02.001
3. Baum S (2023) Assessing natural global catastrophic risks. Nat Hazards 115:2699–2719. https://doi.org/10.1007/s11069-022-05660-w
4. Baum S, Barrett A, Yampolskiy RV (2017) Modeling and interpreting expert disagreement about artificial superintelligence. Informatica 41(7):419–428
5. Beard S, Rowe T, Fox J (2020) An analysis and evaluation of methods currently used to quantify the likelihood of existential hazards. Futures 115:10246. https://doi.org/10.1016/j.futures.2019.102469
6. Betzer A, van den Bongard I, Schweder F et al (2023) All is not lost that is delayed: overconfidence and investment outcomes. Rev Manag Sci 17:2297–2324. https://doi.org/10.1007/s11846-022-00578-w
7. Bostrom N (2002) Existential risks: analyzing human extinction scenarios and related hazards. J Evol Technol 9:1–30
8. Bostrom N (2003) Are we living in a computer simulation? Philos Q 53(211):243–255. https://doi.org/10.1111/1467-9213.00309
9. Buttazzo G (2023) Rise of artificial general intelligence: risks and opportunities. Front Artif Intell. https://doi.org/10.3389/frai.2023.1226990
10. Camerer C, Lovallo D (1999) Overconfidence and excess entry: An experimental approach. Am Econ Rev 89(1):306–318
11. Center for AI Safety (ed) (2023) Mitigating the risk of extinction from AI should be a global priority alongside other societal-scale risks such as pandemics and nuclear war. Center for AI Safety. https://www.safe.ai/work/statement-on-ai-risk. Accessed 29 Aug 2024
12. Charbonneau R (2024) SETI, artificial intelligence, and existential projection. Phys Today 77(2):36–42
13. Cirković MM, Sandberg A, Bostrom N (2010) Anthropic Shadow: Observation Selection Effects and Human Extinction Risks. Risk Anal 30(10):1495–1506. https://doi.org/10.1111/j.1539-6924.2010.01460.x
14. DellaVigna S (2009) Psychology and economics: Evidence from the field. J Econ Lit 47(2):315–72. https://doi.org/10.1257/jel.47.2.315
15. EU (2024) Proposal for a regulation of the european parliament and of the council laying down harmonised rules on artificial intelligence (artificial intelligence act) and amending certain union legislative acts. Legal text, European Union. https://www.google.com/search?client=firefox-bd&q=%3Cdiv+class%3D%22NodiCopyInline%22%3EEU+%282024%29+Proposal+for+a+regulation+of+the+european+parliament+and+of+the+council+laying+down+harmonised+rules+on+a. Accessed 29 Aug 2024

16. Garibay O, Winslow B, Andolina S et al (2023) Six human-centered artificial intelligence grand challenges. Int J Hum Comput Interact 39(3):391–437. https://doi.org/10.1080/10447318.2022.2153320
17. Grace K, Salvatier J, Dafoe A et al (2018) When Will AI Exceed Human Performance? Evidence from AI Experts. J Artif Intell Res 62:729–754
18. Hempsell C (2004) The investigation of natural global catastrophes. J Br Interplanet Soc 57(1/2):2–13
19. Kareiva P, Carranza V (2018) Existential risk due to ecosystem collapse: Nature strikes back. Futures 102:39–50. https://doi.org/10.1016/j.futures.2018.01.001. Futures of research in catastrophic and existential risk
20. Kurzweil R (2005) The Singularity Is Near – When Humans Transcend Biology. Penguin Group, New York
21. Logg JM, Haran U, Moore DA (2018) Is overconfidence a motivated bias? Experimental evidence. J Exp Psychol Gen 147(10):1445
22. McLean S, Read GJM, Thompson J et al (2023) The risks associated with Artificial General Intelligence: A systematic review. J Exp Theor Artif Intell 35(5):649–663. https://doi.org/10.1080/0952813X.2021.1964003
23. Mitchell M (2024) Debates on the nature of artificial general intelligence. Science 383(6689):eado7069. https://doi.org/10.1126/science.ado7069
24. Moravec H (1988) Mind Children – The Future of Robot and Human Intelligence. Harvard University Press, Cambridge, MA
25. Müller VC, Bostrom N (2016) Future progress in artificial intelligence: A survey of expert opinion. In: Müller V (ed) Fundamental Issues of Artificial Intelligence. Springer, pp 553–571
26. Nordhaus W (2021) Are We Approaching an Economic Singularity? Information Technology and the Future of Economic Growth. Am Econ J Macroecon 13(1):299–332. https://doi.org/10.1257/mac.20170105
27. OECD (2012) Mortality risk valuation in environment, health and transport policies. Report, OECD, Paris
28. Rabonza M, Lin Y, Lallemant D (2022) Learning from success, not catastrophe: Using counterfactual analysis to highlight successful disaster risk reduction interventions. Front Earth Sci 10:1–12. https://doi.org/10.3389/feart.2022.847196
29. Rees MJ (2004) Our final century : will civilisation survive the twenty-first century? https://api.semanticscholar.org/CorpusID:191050386. Accessed 29 Aug 2024
30. Robinson LA, Hammitt JK, O'Keeffe L (2019) Valuing mortality risk reductions in global benefit-cost analysis. J Benefit Cost Anal 10(1):15–50. https://doi.org/10.1017/bca.2018.2
31. Rodrik D (2021) Why Does Globalization Fuel Populism? Economics, Culture, and the Rise of Right-Wing Populism. Ann Rev Econ 13:133–170
32. Sagan C (1983) Nuclear war and climatic catastrophe: some policy implications. Foreign Aff; (United States) 62(2):257–292. https://doi.org/10.2307/20041818
33. Schmidt E, Cohen J (2013) The New Digital Age – Transforming Nations, Businesses, and Our Lives. Knopf Doubleday Publishing Group, New York
34. Schubert CLFNSS (2019) The psychology of existential risk: Moral judgments about human extinction. Sci Rep 9(1):15100. https://doi.org/10.1038/s41598-019-50145-9
35. Silver D, Schrittwieser J, Simonyan K (2017) Mastering the game of go without human knowledge. Nature 550:354–359. https://doi.org/10.1038/nature24270
36. Snyder-Beattie A, Ord T, Bonsall M (2019) An upper bound for the background rate of human extinction. Nature Sci Rep 9(11054). https://doi.org/10.1038/s41598-019-47540-7
37. Stokel-Walker C (2022) AI bot ChatGPT writes smart essays-should academics worry? Nature. https://doi.org/10.1038/d41586-022-04397-7
38. Sunstein CR (1999) The law of group polarization. Administrative Law. https://api.semanticscholar.org/CorpusID:145439741. Accessed 29 Aug 2024
39. Sunstein CR (2000) Deliberative trouble - why groups go to extremes. Yale Law J 110:71
40. Sweis N (2022) Revisiting the value of a statistical life: an international approach during covid-19. Risk Manag 24:259–272. https://doi.org/10.1057/s41283-022-00094-x
41. Tetlock, PE, Gardner D (2015) Superforecasting: The Art and Science of Prediction. New York, NY, USA: Crown
42. Tetlock PE, Mellers BA, Scoblic JP (2017) Bringing probability judgments into policy debates via forecasting tournaments. Science 355:481–483
43. Tonn B, MacGregor D (2009) A singular chain of events. Futures 41(10):706–714. https://doi.org/10.1016/j.futures.2009.07.009. Human Extinction
44. Tonn B, Stiefel D (2014) Human extinction risk and uncertainty: Assessing conditions for action. Futures 63:134–144. https://doi.org/10.1016/j.futures.2014.07.001
45. Torres P (2023) Existential Risks: A Philosophical Analysis. Inq Interdiscip J Philos 66(4):614–639. https://doi.org/10.1080/0020174x.2019.1658626
46. Turchin A (2019) Assessing the future plausibility of catastrophically dangerous AI. Futures 107:45–58. https://doi.org/10.1016/j.futures.2018.11.007
47. Turchin A, Denkenberger D (2018) Global catastrophic and existential risks communication scale. Futures 102:27–38. https://doi.org/10.1016/j.futures.2018.01.003. Futures of research in catastrophic and existential risk
48. Viscusi W, Aldy J (2003) The value of a statistical life: A critical review of market estimates throughout the world. J Risk Uncertain 27:5–76. https://doi.org/10.1023/A:1025598106257
49. Weitzman M (2009) On Modeling and Interpreting the Economics of Catastrophic Climate Change. Rev Econ Stat 91(1):1–19. https://doi.org/10.1162/rest.91.1.1
50. Woo G (2018) Counterfactual disadter risk analysis. Variance J 10(2):279–291. Causality Acturial Society
51. Woo G (2021) A counterfactual perspective on compound weather risk. Weather Clim Extremes 32:100314. https://doi.org/10.1016/j.wace.2021.100314
52. Yudkowsky E (2008) Artificial Intelligence as a positive and negative factor in global risk. In Bostrom N, Cirkovic MM (eds) Global Catastrophic Risks, online edn. Oxford, Oxford Academic, 12 Nov 2020. https://doi.org/10.1093/oso/9780198570509.003.0021. Accessed 29 Aug 2024
53. Yudkowsky E (2008) Cognitive biases potentially affecting judgement of global risks. In Bostrom N, Cirkovic MM (eds) Global Catastrophic Risks, online edn. Oxford, Oxford Academic, 12 Nov 2020. https://doi.org/10.1093/oso/9780198570509.003.0009. Accessed 29 Aug 2024

Publisher's Note