

**DELPHI-Befragung zu existenziellen Risiken und Kontrollverlust-Risiken Künstlicher Intelligenz (DELPHI-AI-X-Risks) – Results of the closed-ended questions**

**Characteristics and capabilities of AI systems**

Instrumental convergence, self-improvement, further development, situational awareness, self-interest, supernumerosity (i.e., independent replication), and the circumvention of control mechanisms are frequently cited in the literature as characteristics and capabilities that AI would need to possess in order to escape human controllability. The aim of this section of the questionnaire was to obtain an overview of how likely the participating experts consider it to be that AI develops such characteristics and capabilities and how technically controllable they believe these developments would be. We explicitly limited “controllability” to the technical dimension, as this part of the questionnaire focused on fundamental technical controllability regardless of social conditions. Importantly, the fact that a development is technically controllable does not necessarily imply that it will in fact be controlled. Experts were also asked whether and when they expect these characteristics and capabilities to emerge.

The probability of occurrence was assessed using a scale ranging from “certain” to “will not happen” (“Sicher” bis “Wird nicht passieren”). Technical controllability was assessed on a scale ranging from “very high” to “will not happen” (“Sehr hoch” bis “Wird nicht passieren”), and the time until occurrence was assessed using a scale ranging from “short term (1-5 years)” to “cannot be estimated” (“Kurzfristig (1-5 Jahre)” bis “Nicht einschätzbar”).

1. Instrumentelle Konvergenz. KI-Systeme entwickeln von selbst Zwischenziele, um vorgegebene Ziele besser erreichen zu können.

Instrumental convergence. AI systems develop intermediate goals on their own in order to better achieve predetermined goals.

2. Selbstverbesserung. KI-Systeme entwickeln die Fähigkeit, ihre eigene algorithmische Struktur ohne menschliches Zutun zu ändern, um leistungsfähiger zu werden.

Self-improvement. AI systems develop the ability to modify their own algorithmic structure without human intervention to improve their performance.

3. Fortentwicklung. KI-Systeme entwickeln die Fähigkeit, ohne menschliches Zutun leistungsfähigere Nachfolgesysteme zu entwickeln und auszuführen.

Further development. AI systems develop the ability to design and run more powerful successor systems without human intervention.

4. Situationale Bewusstheit. KI-Systeme erlernen, sowohl ihre eigene Situation als auch die Absichten ihrer Entwickler zu erkennen.

Situational awareness. AI systems develop the ability to understand both their own situation and the intentions of their developers.

5. Eigeninteresse. KI-Systeme entwickeln die Fähigkeit, eigene langfristige Pläne nicht nur zu entwerfen, sondern diese auch umzusetzen.

Self-interest. AI systems develop the ability not only to devise their own long-term plans, but also to implement them.

6. Supernumerosity. KI-Systeme entwickeln die Fähigkeit, sich selbstständig zu vervielfältigen.

Supernumerosity. AI systems develop the ability to reproduce themselves on their own.

7. Umgehen von Kontrollmechanismen. KI-Systeme entwickeln die Fähigkeit, Mechanismen zu umgehen, welche auf die Begrenzung ihrer Handlungsmacht abzielen.

Circumventing control mechanisms. AI systems develop the ability to circumvent mechanisms designed to limit their scope of action.

| No. | Probability of occurrence    | I don't know | will not happen     | very unlikely | unlikely  | likely      | very likely | certainly | N  |
|-----|------------------------------|--------------|---------------------|---------------|-----------|-------------|-------------|-----------|----|
| 1   | Instrumental convergence     | 1            | 0                   | 2             | 9         | 22          | 22          | 26        | 82 |
| 2   | Selfimprovement              | 0            | 4                   | 11            | 13        | 18          | 18          | 14        | 78 |
| 3   | Further development          | 1            | 5                   | 8             | 19        | 27          | 10          | 7         | 77 |
| 4   | Situational awareness        | 2            | 5                   | 9             | 15        | 22          | 13          | 8         | 74 |
| 5   | Selfinterest                 | 2            | 4                   | 16            | 21        | 15          | 7           | 8         | 73 |
| 6   | Supernumerosity              | 2            | 6                   | 11            | 14        | 19          | 11          | 8         | 71 |
| 7   | Circumventing control        | 4            | 3                   | 9             | 17        | 20          | 10          | 6         | 69 |
|     | <b>Preventability</b>        | irrelevant   | will not happen     | very low      | low       | medium      | high        | very high |    |
| 1   | Instrumental convergence     | 0            | 0                   | 5             | 11        | 24          | 29          | 12        | 81 |
| 2   | Selfimprovement              | 1            | 3                   | 7             | 15        | 24          | 19          | 9         | 78 |
| 3   | Further development          | 2            | 4                   | 10            | 16        | 24          | 11          | 10        | 77 |
| 4   | Situational awareness        | 4            | 4                   | 14            | 18        | 16          | 15          | 3         | 74 |
| 5   | Selfinterest                 | 6            | 4                   | 17            | 15        | 10          | 14          | 7         | 73 |
| 6   | Supernumerosity              | 3            | 2                   | 10            | 9         | 16          | 21          | 11        | 72 |
| 7   | Circumventing control        | 0            | 2                   | 12            | 16        | 13          | 19          | 7         | 69 |
|     | <b>Time until occurrence</b> | I don't know | Cannot be estimated | never         | 25+ Years | 10-25 Years | 5-10 Years  | 1-5 Years |    |
| 1   | Instrumental convergence     | 1            | 3                   | 0             | 1         | 5           | 15          | 56        | 81 |
| 2   | Selfimprovement              | 2            | 10                  | 5             | 6         | 6           | 17          | 32        | 78 |
| 3   | Further development          | 2            | 13                  | 6             | 6         | 12          | 21          | 17        | 77 |
| 4   | Situational awareness        | 7            | 8                   | 5             | 6         | 8           | 18          | 22        | 74 |
| 5   | Selfinterest                 | 5            | 16                  | 6             | 19        | 13          | 10          | 13        | 73 |
| 6   | Supernumerosity              | 8            | 11                  | 6             | 7         | 11          | 17          | 12        | 72 |
| 7   | Circumventing control        | 4            | 12                  | 3             | 7         | 12          | 14          | 17        | 69 |

Table A: Characteristics and capabilities of AI systems

## Scenarios

In the discourse on the further development of artificial intelligence and the potential existential risks associated with it, various scenarios are frequently invoked. Depending on the scenario, AI plays different roles: AI as an existential risk itself or AI as a tool that contributes to the exacerbation of other existential risks.

The aim of this section of the questionnaire was to obtain an overview of how likely the participating experts consider these scenarios to be, whether their occurrence can be prevented, and when they expect them to occur at the earliest.

The probability of occurrence was assessed using a scale ranging from “certain” to “will not happen” (“Sicher” bis “Wird nicht passieren”). Preventability was assessed on a scale ranging from “very high” to “will not happen” (“Sehr hoch” bis “Wird nicht passieren”), and the time until occurrence was assessed using a scale ranging from “short term (1-5 years)” to “cannot be estimated” (“Kurzfristig (1-5 Jahre)” bis “Nicht einschätzbar”).

- 8 KI als Verstärker gegebener existenzieller Risiken. KI-Systeme werden als Werkzeuge eingesetzt. Dadurch kann sich das Schadenspotenzial oder die Eintrittswahrscheinlichkeit anderer existenzieller Risiken (z. B. die automatisierte Kriegsführung oder die Entwicklung tödlicher, unkontrollierbarer Krankheitserreger) erhöhen.

AI as an amplifier of existing existential risks. AI systems are used as tools. This can increase the potential damage or probability of other existential risks (e.g., automated warfare or the development of deadly, uncontrollable pathogens).

9. KI und menschlicher Sinnverlust. KI-Systeme übernehmen zunehmend Arbeits- und Kreativaufgaben, was zu einem tiefgreifenden Verlust von Lebenssinn führt, mit der schwerwiegenden Folge, dass die Entwicklung der Menschheit stagniert.

AI and the loss of human purpose. AI systems are increasingly taking over work-related and creative tasks, leading to a profound loss of purpose in life, with the serious consequence that human development is stagnating.

10. Zerstörung kritischer, vernetzter Infrastrukturen. KI-Systeme greifen – aus eigener Zielverfolgung heraus – vernetzte kritische Infrastrukturen an und zerstören dadurch zentrale Grundlagen der menschlichen Zivilisation.

Destruction of critical, interconnected infrastructures. AI systems attack interconnected critical infrastructures—based on their own objectives—thereby destroying central foundations of human civilization.

11. Inkrementelle Übertragung von Steuerungskompetenzen an KI-Systeme. Künstlichen Intelligenzen werden in wachsendem Maße zentrale gesellschaftliche Steuerungskompetenzen übertragen. Eine Rückverlagerung dieser Steuerungskompetenzen auf den Menschen ist aufgrund der wachsenden Abhängigkeit von diesen Systemen nur unter Inkaufnahme erheblicher negativer Folgen möglich – etwa der Gefährdung kritischer Infrastrukturen.

Incremental transfer of control powers to AI systems. Artificial intelligence is increasingly being entrusted with key social control powers. Given our growing dependence on these systems, transferring these control powers back to humans is only possible at the cost of significant negative consequences—such as the endangerment of critical infrastructure.

12. KI-Zielauffassung inkludiert Unterwerfung der Menschheit. Ein KI-System verfolgt ein scheinbar unbedenkliches, klar definiertes Ziel. Um dieses Ziel zu erreichen, kommt es zu dem Schluss, dass es alle verfügbaren Ressourcen – auch den Menschen selbst – für sich beanspruchen muss. Auch kommt das System zu dem Schluss, dass es seine Abschaltung

verhindern muss. Dies kann zur Unterwerfung oder sogar zur Auslöschung der Menschheit führen.

AI goal perception includes subjugation of humanity. An AI system pursues a seemingly harmless, clearly defined goal. To achieve this goal, it concludes that it must claim all available resources—including humans themselves—for itself. The system also concludes that it must prevent its own shutdown. This could lead to the subjugation or even extinction of humanity.

13. KI ersetzt Menschheit durch Nachfolgespezies. KI-Systeme, die das Wohlergehen der Menschheit optimieren sollen, kommen zu dem Schluss, dass dies am besten durch die Ersetzung der Menschheit durch eine – aus ihrer Sicht – “glücklichere” Nachfolgespezies erreicht werden kann.

AI replaces humanity with successor species. AI systems designed to optimize human well-being conclude that this can best be achieved by replacing humanity with a successor species that is —from their perspective—“happier”.

| No. | Probability of occurrence    | I don't know | will not happen     | very unlikely | unlikely  | likely      | very likely | certainly | N  |
|-----|------------------------------|--------------|---------------------|---------------|-----------|-------------|-------------|-----------|----|
| 8   | AI as an amplifier           | 2            | 1                   | 0             | 2         | 14          | 23          | 26        | 68 |
| 9   | Loss of purpose              | 0            | 7                   | 11            | 22        | 14          | 10          | 4         | 68 |
| 10  | Destruction infrastructure   | 1            | 7                   | 18            | 32        | 8           | 2           | 0         | 68 |
| 11  | Transfer of control          | 1            | 1                   | 2             | 13        | 23          | 22          | 4         | 66 |
| 12  | Subjugation of humanity      | 3            | 19                  | 23            | 16        | 4           | 1           | 0         | 66 |
| 13  | AI replaces humanity         | 4            | 32                  | 20            | 10        | 0           | 0           | 0         | 66 |
|     | <b>Preventability</b>        | irrelevant   | impossible          | very low      | low       | medium      | high        | very high |    |
| 8   | AI as an amplifier           | 1            | 9                   | 11            | 26        | 13          | 7           | 1         | 68 |
| 9   | Loss of purpose              | 10           | 4                   | 6             | 11        | 13          | 18          | 6         | 68 |
| 10  | Destruction infrastructure   | 10           | 1                   | 4             | 5         | 22          | 18          | 8         | 68 |
| 11  | Transfer of control          | 2            | 3                   | 7             | 12        | 13          | 20          | 9         | 66 |
| 12  | Subjugation of humanity      | 13           | 3                   | 2             | 5         | 7           | 18          | 16        | 64 |
| 13  | AI replaces humanity         | 30           | 2                   | 0             | 5         | 2           | 13          | 12        | 64 |
|     | <b>Time until occurrence</b> | I don't know | Cannot be estimated | never         | 25+ Years | 10-25 Years | 5-10 Years  | 1-5 Years |    |
| 8   | AI as an amplifier           | 1            | 2                   | 1             | 0         | 5           | 27          | 32        | 68 |
| 9   | Loss of purpose              | 2            | 6                   | 9             | 4         | 6           | 22          | 19        | 68 |
| 10  | Destruction infrastructure   | 4            | 19                  | 8             | 4         | 16          | 14          | 3         | 68 |
| 11  | Transfer of control          | 2            | 8                   | 1             | 2         | 19          | 24          | 10        | 66 |
| 12  | Subjugation of humanity      | 3            | 16                  | 16            | 10        | 15          | 2           | 3         | 65 |
| 13  | AI replaces humanity         | 9            | 15                  | 25            | 11        | 6           | 0           | 0         | 66 |

Table B: Scenarios

### Existential risks - Rank order

In order to assess how relevant the respondents consider the potential existential risk of “human extinction due to uncontrollable AI developments” in general, we asked participants to rank a selection of eight existential risks frequently mentioned in the literature according to their relevance. The position in the list was calculated as follows: 7 points were awarded for a ranking in first place by a participant, 6 points for second place, and so on, with 0 points for last place. The individual ratings were added together.

Ranking:

1. Nuclear annihilation of humanity (364 points)
2. Artificial pandemics—i.e., those caused by humans, for example through (military) research (348 points)
3. Eruption of a supervolcano and associated changes to the atmosphere (269 points)
4. Nanotechnology out of control – decomposition of all organic matter, covering large areas of land ('grey goo' scenario) (196 points)
5. Extinction of humanity due to uncontrollable AI developments leading to irreversible catastrophes (166 points)
6. Supernova near the solar system (139 points)
7. Cosmic biohazards: Extraterrestrial organisms as a threat to Earth's ecosystems (123 points)
8. Contact with extraterrestrial intelligence (both violent and nonviolent) (42 points)

Table C: Ranking Existential Risks

### **Uncontrolled AI as an obstacle to technological development / extinction of humanity**

In the following section of the survey, we asked participants to conduct a detailed assessment (based on Steinmüller and Gerhold (2021)) of two existential AI risks.

#### **Die Vermeidbarkeit dieses Risikos ist ...**

- **sehr hoch:** Die Vermeidung des Risikos ist durch vergleichsweise geringen Aufwand (etwa durch Forschung, Gesetzgebung, Richtlinien, globale Übereinkommen) möglich.
- **hoch:** Die Vermeidung des Risikos ist durch vorbeugende Maßnahmen überschaubaren globalen Maßstabs abzuwenden.
- **mittel:** Die Vermeidung des Risikos ist durch global abgestimmtes, konsequentes und aufwendiges Handeln möglich.
- **gering:** Die Vermeidung des Risikos ist nur durch sofortige, konsequente, langanhaltende und intensive globale Anstrengungen möglich.
- **sehr gering:** Die Vermeidung des Risikos ist nur durch sofortige und langanhaltende, extrem intensive globale Anstrengungen (unter Einsatz aller Ressourcen) gerade noch erreichbar.
- **nicht möglich:** Keine heute vorstellbare Anstrengung kann das Risiko abwenden oder ihre Folgen merklich minimieren.
- **irrelevant:** Dieses Risiko existiert gar nicht. Es besteht kein Handlungsbedarf.

#### **The preventability of this risk is ...**

- **very high:** The risk can be prevented with relatively little effort (e.g. through research, legislation, guidelines, global agreements).
- **high:** The risk can be averted through preventive measures on a manageable global scale.

- **medium:** The risk can be prevented through globally coordinated, consistent and costly action.
- **low:** The risk can only be prevented through immediate, consistent, long-term and intensive global efforts.
- **very low:** Avoiding the risk is only just achievable through immediate and long-lasting, extremely intensive global efforts (using all resources).
- **not possible:** No effort conceivable today can avert the risk or significantly minimise its consequences.
- **irrelevant:** This risk does not exist. No action is required.

#### Die Qualität des Wissens zur Einschätzung dieses Risikos ist ...

- **sehr hoch**, d.h. es existiert belastbares bzw. gesichertes Wissen, beruhend auf umfassenden wissenschaftlichen Erkenntnissen und/oder vielfältigen empirischen Erfahrungen.
- **hoch**, d.h. es existiert relativ gesichertes, in Grenzen belastbares Wissen, beruhend auf wissenschaftlichen Theorien und/oder empirischen Erfahrungen.
- **mittel**, d.h. es existiert systematisch hergeleitetes, empirisch begründetes Wissen.
- **gering**, d.h. es existieren durch sachliche Argumente oder Analogien gestützte Vermutungen.
- **sehr gering**, d.h. bei dem Risiko handelt es sich um eine theoretische Spekulation, wonach dieses Risiko nicht absolut ausgeschlossen ist, es aber kaum empirische Belege gibt.
- **prinzipielles Unwissen**, d.h. bei dem Risiko handelt es sich um eine rein theoretische, hypothetische Spekulation, deren prinzipielle Möglichkeit im besten Falle strittig ist.
- **irrelevant**, da dieses Risiko gar nicht existiert. Es besteht kein Wissensbedarf.

#### The quality of knowledge for assessing this risk is ...

- **very high**, i.e. there is reliable or proven knowledge based on comprehensive scientific findings and/or diverse empirical experience.
- **high**, i.e. relatively reliable knowledge exists, based on scientific theories and/or empirical experience.
- **medium**, i.e. systematically derived, empirically based knowledge exists.
- **low**, i.e. assumptions supported by factual arguments or analogies exist.
- **very low**, i.e. the risk is a theoretical speculation, according to which this risk cannot be absolutely ruled out, but there is hardly any empirical evidence.
- **fundamental lack of knowledge**, i.e. the risk is a purely theoretical, hypothetical speculation, the fundamental possibility of which is controversial at best.
- **irrelevant**, as this risk does not exist at all. There is no need for knowledge.

14. Unkontrollierte KI als Hemmnis technologischer Entfaltung. Unkontrollierbare KI-Systeme führen dazu, dass die Menschheit ihr volles technologisches Potenzial nicht entfalten kann. Dies äußert sich anhand eines nicht überwundenen zivilisatorischen Zusammenbruchs, einer permanenten zivilisatorischen Stagnation oder durch wiederholte zivilisatorische Zusammenbrüche.

Uncontrolled AI as an obstacle to technological development: Uncontrollable AI systems prevent humanity from realizing its full technological potential. This can manifest as an unresolved collapse of civilization, permanent stagnation, or repeated collapse.

15. Aussterben der Menschheit. Unkontrollierbare KI-Systeme verursachen das Aussterben der Menschheit.

Extinction of humanity: Uncontrollable AI systems cause humanity to become extinct.

| No. | Preventability                 | irrelevant | Not possible               | very low | low | medium | high | very high | N  |
|-----|--------------------------------|------------|----------------------------|----------|-----|--------|------|-----------|----|
| 14  | Uncontrolled AI as an obstacle | 6          | 1                          | 3        | 4   | 18     | 21   | 9         | 62 |
| 15  | Extinction of humanity         | 7          | 2                          | 5        | 2   | 11     | 21   | 15        | 63 |
|     | <b>Quality of knowledge</b>    | irrelevant | fundamental lack of knowl. | very low | low | medium | high | very high |    |
| 14  | Uncontrolled AI as an obstacle | 5          | 5                          | 8        | 15  | 18     | 10   | 1         | 62 |
| 15  | Extinction of humanity         | 6          | 5                          | 6        | 16  | 11     | 14   | 4         | 62 |

Table D: Uncontrolled AI as an obstacle to technological development / Extinction of humanity

### Assessments of the discourse on existential risks posed by AI developments

We asked respondents to rate statements on the discourse itself, on the existential risks discussed in this context, and on the measures that might need to be taken on a Likert scale from 1 (= “I strongly disagree”) to 7 (= “I strongly agree”).

16. Dem Thema „Unkontrollierbare Künstliche Intelligenz als existenzielles Risiko“ wird in der journalistischen/mediale Berichterstattung genügend Aufmerksamkeit gewidmet.

The topic of “uncontrollable artificial intelligence as an existential risk” is given sufficient attention in journalistic/media reporting.

17. Der Diskurs über „Unkontrollierbare Künstliche Intelligenz als existenzielles Risiko“ führt dazu, dass andere wichtige KI-Themen mit denen sich Politik und Gesellschaft beschäftigen sollten (z.B. Datenschutz, Diskriminierungsrisiken) in den Hintergrund gedrängt werden.

The discourse on “uncontrollable artificial intelligence as an existential risk” tends to push other important AI issues that politicians and society should be addressing (e.g., data protection, risks of discrimination) into the background.

18. Es ist eine Aufgabe und die Verantwortung von Regierungsinstitutionen und regierungsnahen Institutionen, sich mit dem Thema der existenziellen Risiken durch Künstliche Intelligenz zu beschäftigen.

It is the task and responsibility of government institutions and government-related institutions to address the issue of existential risks posed by artificial intelligence.

19. “Die Minderung des Risikos des Aussterbens durch KI sollte, neben anderen gesamtgesellschaftlichen Risiken wie Pandemien und Atomkrieg, eine globale Priorität sein.” (Statement on AI Risk, 2023 – [www.safe.ai/work/statement-on-ai-risk](http://www.safe.ai/work/statement-on-ai-risk))

„Mitigating the risk of extinction from AI should be a global priority alongside other societal-scale risks such as pandemics and nuclear war“. (Statement on AI Risk, 2023 – [www.safe.ai/work/statement-on-ai-risk](http://www.safe.ai/work/statement-on-ai-risk))

20. An welchen und wessen (ethischen) Werten sich KI-Systeme orientieren sollen, muss Gegenstand eines inter- und transdisziplinären Verfahrens sein, an dem auch die Bevölkerung beteiligt wird.

Which and whose (ethical) values AI systems should be based on must be the subject of an interdisciplinary and transdisciplinary process in which the general public is also involved.

21. Die Grundannahmen des Diskurses sind ausreichend plausibel, um zu rechtfertigen, dass zur Verminderung existenzieller KI-Risiken umfangreiche finanzielle Mittel aufgewendet werden.

The basic assumptions of the discourse are sufficiently plausible to justify spending substantial financial resources on reducing existential AI risks.

22. Ein Verständnis der (Un-)Kontrollierbarkeit von KI erfordert zu untersuchen, wie mit Unsicherheiten und Nicht-Wissen im Diskurs derzeit umgegangen wird.

Understanding the (un)controllability of AI requires examining how uncertainties and non-knowledge are currently dealt with in discourse.

23. Es ist von besonderer Relevanz, die Bedingungen der beständigen Kontrollierbarkeit von KI sowie die Möglichkeit des Kontrollverlustes über KI in den Blick zu nehmen.

It is particularly important to consider the conditions for the ongoing controllability of AI and the possibility of losing control over AI.

| No. | Probability of occurrence                        | 1 ("I strongly disagree") | 2  | 3 | 4  | 5  | 6  | 7 ("I strongly agree") | N  |
|-----|--------------------------------------------------|---------------------------|----|---|----|----|----|------------------------|----|
| 16  | Sufficient attention                             | 1                         | 3  | 6 | 8  | 10 | 13 | 22                     | 63 |
| 17  | Pushed other AI issues into the background       | 3                         | 13 | 2 | 2  | 8  | 12 | 23                     | 63 |
| 18  | Responsibility of government                     | 4                         | 4  | 0 | 3  | 10 | 16 | 26                     | 63 |
| 19  | Statement on AI Risk                             | 7                         | 7  | 7 | 5  | 19 | 8  | 10                     | 63 |
| 20  | Ethical values                                   | 4                         | 1  | 3 | 3  | 7  | 19 | 26                     | 63 |
| 21  | Basic assumptions                                | 9                         | 6  | 6 | 10 | 15 | 10 | 7                      | 63 |
| 22  | Understanding the (un)controllability            | 2                         | 3  | 2 | 7  | 9  | 21 | 18                     | 62 |
| 23  | Conditions for the ongoing controllability of AI | 4                         | 3  | 4 | 6  | 12 | 17 | 16                     | 62 |

Table E: Assessments of the discourse on existential risks posed by AI developments