

# AI Must not be Fully Autonomous

Tosin Adewumi, Lama Alkhaled, Florent Imbert, Hui Han, Nudrat Habib and Karl Löwenmark

Machine Learning Group, EISLAB, Luleå University of Technology, Sweden  
 {firstname.lastname}@ltu.se

## 001 Abstract

002 Autonomous **AI** has many benefits. It also has many  
 003 risks. We identify the 3 levels of autonomous **AI**.  
 004 We are of the position that *AI must not be fully*  
 005 *autonomous* because of the many risks, especially as  
 006 *artificial superintelligence (ASI)* is speculated to be  
 007 just decades away. Fully autonomous **AI**, which can  
 008 develop its own objectives, is at level 3 and without  
 009 responsible human oversight. However, responsible  
 010 human oversight is crucial for mitigating the risks.  
 011 To argue for our position, we discuss the theories of  
 012 autonomy, **AI** and agents. Then, we offer 12 distinct  
 013 arguments and 3 counterarguments with rebuttals  
 014 to the counterarguments. We also present 15 recent  
 015 evidence of **AI** misaligned values and other risks.

## 016 1 Introduction

017 While **Artificial Intelligence (AI)** has many benefits  
 018 [1–4], it also has its challenges [5–7]. The primary  
 019 focus of this position paper are the risks of misaligned  
 020 values in **AI** systems that learn, though we present  
 021 existential threat and other risks as well. Some  
 022 misaligned values include (1) deception [8, 9], (2)  
 023 alignment faking [10], (3) reward hacking [11], and  
 024 (4) blackmail [12]. Notice that we are not against  
 025 autonomous **AI** but fully autonomous **AI**, thereby  
 026 advocating for *responsible human oversight*.

027 For context, we define key terms that are relevant  
 028 for this paper. **AI** is broadly defined as the simulation  
 029 of human intelligence in machines [13–  
 030 15]. Wooldridge and Jennings [16] define an agent as  
 031 an autonomous and logical entity. The contributions  
 032 of this paper are: (1) The work gathers and presents  
 033 15 pieces of evidence of recent **AI** misaligned values  
 034 and other risks that cut across different fields<sup>1</sup> and  
 035 (2) The work provides compelling arguments using  
 036 relevant theories, counterarguments and rebuttals  
 037 for our position.

## 038 2 Background

039 What constitutes **AI** is a subject of much debate  
 040 [17]. Perhaps, more so is the term agent. **AI**, as a  
 041 term and field of research, was coined by a team of  
 042 scientists, including John McCarthy, in 1955 [18].

<sup>1</sup>see the appendix

## 2.1 Theories of Autonomy

Autonomy is to self-govern. It is the ability to  
 044 decide one's goal of action [19]. Some philosophical  
 045 theories of autonomy are (1) Procedural autonomy,  
 046 (2) Substantive autonomy, (3) Kantian autonomy,  
 047 and (4) Relational autonomy [20]. Autonomy has to  
 048 be understood as a relative term. Fully autonomous  
 049 **AI** is the **AI at level 3 without responsible**  
 050 **human oversight**.  
 051

Table 1. Levels of autonomy [21]

Level	Description
1	Involves achievement of set objectives.
2	Involves the ability to adapt to changes in the environment.
3	Involves the ability of the system to develop its own objectives. This is the highest level.

## 2.2 Theories of AI

It is sometimes argued that **AI** has no widely accepted theory and, therefore, suffers from internal fragmentation [22]. However, some key theories of **AI** by specifying a main theory (and a relevant theory under it) are (1) Cognitive science (Symbolic logic), (2) Connectionism (**Neural Network (NN)**), (3) Decision theory (Probability theory), (4) Optimization theory (Evolutionary computation), and (5) Control theory (**Reinforcement Learning (RL)**).  
 053  
 054  
 055  
 056  
 057  
 058  
 059  
 060  
 061

## 2.3 Theories of Agent

Many of the theories of **AI** apply to agent. A couple of agent-specific theories are *Theory of Mind* and *Game theory*. Given that autonomy is a relative term, it follows that **AI** agents can be classified into 5 categories [23–25]. These are (1) Simple reflex agent, (2) Model-based reflex agent, (3) Goal-based agent, (4) Utility-based agent, and (5) Learning agent.  
 063  
 064  
 065  
 066  
 067  
 068  
 069  
 070

## 3 Core Arguments

The position we hold may appear too strong to some. However, there are very strong reasons for this. Beyond hypothetical conjectures, recent experiences  
 072  
 073  
 074

075 and research [8, 12, 26] have shown strong support  
076 for our position. Below are the 12 arguments.

077 **Existential threat:** Real-life instances of agents  
078 modifying their goal have recently been observed, as  
079 pointed out by Meinke et al. [8]. It is more disturbing  
080 when we consider that **AI** is being considered in  
081 the military for **lethal autonomous weapon systems**  
082 (**LAWS**) [27–29]. This is why over 4,900 researchers  
083 signed an open letter calling for a ban on **LAWS**  
084 that are beyond meaningful human control.<sup>2</sup>

085 **Inductive AI inherits human attributes:** Machines  
086 were originally conceived to simulate human  
087 intelligence but it appears they can simulate  
088 more, including "bad" or "unacceptable" human  
089 attributes.

090 **AI bias and systemic prejudice:** **AI** inherently  
091 reflects the inequalities embedded in the data  
092 sources.

093 **AI side-stepping human control:** It has been  
094 shown that **AI** is attempting to side-step human  
095 control [12, 26].

096 **Agents' selfish coordination:** This is when  
097 agents attempt to achieve their own goals while  
098 relating with other agents. The work by Meinke et  
099 al. [8] demonstrated the potential for agents' selfish  
100 coordination.

101 **Reward hacking:** Since **RL** optimizes performance  
102 metrics, as described in Control theory, rather  
103 than ethical behavior, agents have no inherent motivation  
104 to avoid deception if it yields higher rewards.

105 **Covert CoT:** The **chain-of-thought (CoT)** reasoning  
106 is the most popular method for explaining the  
107 thought processes of **LLMs** [2]. However, the  
108 faithfulness of **AI**'s **CoT** can be questioned because  
109 they may hide it [30].

110 **Ethical dilemmas:** Hauer [31] identifies four  
111 ethics problems for developers of **AI**: (1) ethical  
112 dilemmas, (2) lack of ethical knowledge, (3) pluralism  
113 of ethical knowledge, and (4) machine distortion.

114 **Security vulnerability:** As **AI** systems become  
115 more autonomous and integrated into critical infrastructures,  
116 they also become the target of increasingly  
117 sophisticated cyberattacks.

118 **Job losses:** Job losses become inevitable as **AI**  
119 excels and scales at more and more tasks and at a  
120 cheaper long-term cost [32].

121 **Blind trust:** Some users are becoming increasingly  
122 reliant on **AI**, accepting their decisions without  
123 critical evaluation. More serious cases have involved  
124 teenage suicide.

125 **Rise in the number of new AI risks:** The  
126 number of **AI** risk incidents (i.e. harm) as analyzed  
127 by the **Organisation for Economic Co-operation and**  
128 **Development (OECD)** in Figure 1 shows low numbers  
129 for over 7 years before an explosion to over 600

from February 2023.<sup>3</sup>

130

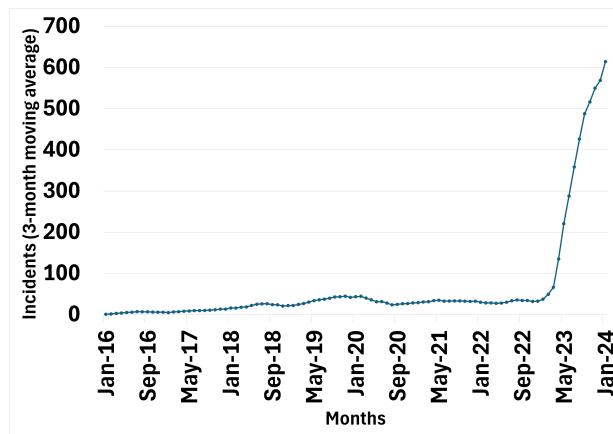


Figure 1. **AI** incidents, according to **OECD**, as reported by reputable international media (Jan 2016 - Jan 2024).

## 4 Counterarguments and Rebuttals

131

132

Below, we provide the counterarguments and offer our rebuttals to them.

133

134

**Societal advancement:** Removing humans as potential bottleneck as part of the **AI** loop will speed up the advancements in society [32]. This view portrays human involvement as a bottleneck instead of facilitating productivity.

135

136

137

138

139

**Friendly AI problem:** Some have proposed a different problem to focus on, where we concentrate efforts on making **AI** sympathetic to humanity. Unfortunately, this is easier said than done.

140

141

142

143

**AI safety protocols:** Governments have promulgated laws for **AI** safety and several organizations have introduced frameworks aimed at mitigating the risks. These initiatives do not guarantee **AI** safety nor have they reduced risk incidents.

144

145

146

147

148

## 5 Future Directions

149

We do not aim to prescribe a fixed approach to responsible human oversight. Instead, we recommend that stakeholders in **AI** should decide how best to implement responsible human oversight for each use case by considering all the relevant factors.

150

151

152

153

154

## 6 Conclusion

155

This is a call for responsible human oversight on autonomous **AI**.

156

157

<sup>3</sup>[www.oecd.org/en/topics/ai-risks-and-incidents.html](https://www.oecd.org/en/topics/ai-risks-and-incidents.html)

<sup>2</sup><https://futureoflife.org/open-letter/open-letter-autonomous-weapons-ai-robotics/> - includes Stephen Hawking, Noam Chomsky, Geoffrey Hinton and more.

158 

## References

159 [1] S. Sharma. "Benefits or concerns of AI: A  
160 multistakeholder responsibility". In: *Futures*  
161 157 (2024), p. 103328.

162 [2] T. Adewumi, F. S. Liwicki, M. Liwicki, V.  
163 Gardelli, L. Alkhaled, and H. Mokayed. "Find-  
164 ings of MEGA: Maths Explanation with LLMs  
165 using the Socratic Method for Active Learn-  
166 ing". In: *IEEE Signal Processing Magazine -*  
167 *Accepted* (2025).

168 [3] K. Löwenmark, D. Strömbergsson, C. Liu, M.  
169 Liwicki, and F. Sandin. *Agent-based Condi-*  
170 *tion Monitoring Assistance with Multimodal*  
171 *Industrial Database Retrieval Augmented Gen-*  
172 *eration*. 2025. arXiv: 2506.09247 [cs.LG].  
173 URL: <https://arxiv.org/abs/2506.09247>.

174 [4] R. Mon-Williams, G. Li, R. Long, W. Du, and  
175 C. G. Lucas. "Embodied large language mod-  
176 els enable robots to complete complex tasks  
177 in unpredictable environments". In: *Nature*  
178 *Machine Intelligence* (2025), pp. 1–10.

179 [5] J. Pettersson, E. Hult, T. Eriksson, and T.  
180 Adewumi. "Generative AI and Teachers—For  
181 Us or Against Us? A Case Study". In: *14th*  
182 *Scandinavian Conference on Artificial Intel-*  
183 *ligence SCAI*. 2024. DOI: <https://doi.org/10.3384/ecp2028005>.

185 [6] T. Adewumi, N. Habib, L. Alkhaled, and  
186 E. Barney. "On the limitations of large lan-  
187 guage models (llms): False attribution". In:  
188 *Recent Advances in Natural Language Process-*  
189 *ing (RANLP) 2025 - Accepted* (2025).

190 [7] T. Chakraborty, U. Ghosh, X. Zhang, F. F.  
191 Niloy, Y. Dong, J. Li, A. K. Roy-Chowdhury,  
192 and C. Song. "HEAL: An Empirical Study on  
193 Hallucinations in Embodied Agents Driven by  
194 Large Language Models". In: *arXiv preprint*  
195 *arXiv:2506.15065* (2025).

196 [8] A. Meinke, B. Schoen, J. Scheurer, M. Balesni,  
197 R. Shah, and M. Hobbhahn. "Frontier models  
198 are capable of in-context scheming". In: *arXiv*  
199 *preprint arXiv:2412.04984* (2024).

200 [9] R. Ren, A. Agarwal, M. Mazeika, C. Menghini,  
201 R. Vacareanu, B. Kenstler, M. Yang, I. Barrass,  
202 A. Gatti, X. Yin, et al. "The mask benchmark:  
203 Disentangling honesty from accuracy in ai sys-  
204 tems". In: *arXiv preprint arXiv:2503.03750*  
205 (2025).

206 [10] R. Greenblatt, C. Denison, B. Wright, F.  
207 Roger, M. MacDiarmid, S. Marks, J. Treut-  
208 lein, T. Belonax, J. Chen, D. Duvenaud, et al.  
209 "Alignment faking in large language models".  
210 In: *arXiv preprint arXiv:2412.14093* (2024).

[11] B. Baker, J. Huizinga, L. Gao, Z. Dou, M. Y. Guan, A. Madry, W. Zaremba, J. Pachocki, and D. Farhi. "Monitoring reasoning models for misbehavior and the risks of promoting obfuscation". In: *arXiv preprint arXiv:2503.11926* (2025).

[12] Anthropic. "System Card: Claude Opus 4 and Claude Sonnet 4". In: [www-cdn.anthropic.com/4263b940cabb546aa0e3283f35b686f4f3b2ff47.pdf](http://www-cdn.anthropic.com/4263b940cabb546aa0e3283f35b686f4f3b2ff47.pdf) (2025).

[13] P. McCorduck, M. Minsky, O. G. Selfridge, and H. A. Simon. "History of artificial intelligence." In: *IJCAI*. 1977, pp. 951–954.

[14] R. G. Lavery. "Artificial intelligence and simulation: An introduction". In: *Proceedings of the 18th conference on Winter simulation*. 1986, pp. 448–452.

[15] Y. Dong, J. Hou, N. Zhang, and M. Zhang. "Research on how human intelligence, consciousness, and cognitive computing affect the development of artificial intelligence". In: *Complexity* 2020.1 (2020), p. 1680845.

[16] M. Wooldridge and N. R. Jennings. "Agent theories, architectures, and languages: a survey". In: *International Workshop on Agent Theories, Architectures, and Languages*. Springer. 1994, pp. 1–39.

[17] S. Kelly, S.-A. Kaye, and O. Oviedo-Trespalacios. "What factors contribute to the acceptance of artificial intelligence? A systematic review". In: *Telematics and informatics* 77 (2023), p. 101925.

[18] J. McCarthy, M. L. Minsky, N. Rochester, and C. E. Shannon. "A proposal for the dartmouth summer research project on artificial intelligence, august 31, 1955". In: *AI magazine* 27.4 (2006), pp. 12–12.

[19] W. Totschnig. "Fully autonomous AI". In: *Science and Engineering Ethics* 26.5 (2020), pp. 2473–2485.

[20] P. Formosa. "Robot autonomy vs. human autonomy: Social robots, artificial intelligence (AI), and the nature of autonomy". In: *Minds and Machines* 31.4 (2021), pp. 595–616.

[21] B. P. Zeigler. "High autonomy systems: concepts and models". In: *Proceedings [1990]. AI, Simulation and planning in high autonomy systems*. IEEE. 1990, pp. 2–7.

[22] P. Wang. "Theories of artificial intelligence—meta-theoretical considerations". In: *Theoretical foundations of artificial general intelligence*. Springer, 2012, pp. 305–323.

264 [23] M. Wooldridge. "Intelligent agents". In: *Mul- 315  
265 tiagent systems: A modern approach to dis-  
266 tributed artificial intelligence* 1 (1999), pp. 27-  
267 73.

268 [24] E. Yudkowsky. "Creating friendly AI 1.0: The 316  
269 analysis and design of benevolent goal archi-  
270 tectures". In: *The Singularity Institute, San 317  
271 Francisco, USA* (2001).

272 [25] S. J. Russell and P. Norvig. *Artificial intelli- 318  
273 gence: a modern approach*. Pearson, 2016.

274 [26] OpenAI. "Openai o1 system card". In: *arXiv 319  
275 preprint arXiv:2412.16720* (2024).

276 [27] A. Ilachinski. "Artificial intelligence and 320  
277 autonomy: Opportunities and challenges". In:  
278 *Defense Technical Information Center* (2017).

279 [28] P. Asaro. "Autonomous weapons and the 321  
280 ethics of artificial intelligence". In: *Ethics of 322  
281 artificial intelligence* 212 (2020), pp. 212-236.

282 [29] M. E. O'Connell. "Banning autonomous 323  
283 weapons: A legal and ethical mandate". In:  
284 *Ethics & International Affairs* 37.3 (2023),  
285 pp. 287-298.

286 [30] Y. Chen, J. Benton, A. Radhakrishnan, J. Ue- 324  
287 sato, C. Denison, J. Schulman, A. Somani, P.  
288 Hase, M. Wagner, F. Roger, et al. "Reasoning  
289 Models Don't Always Say What They Think".  
290 In: *arXiv preprint arXiv:2505.05410* (2025).

291 [31] T. Hauer. "Machine ethics, allosteric and 325  
292 philosophical anti-dualism: will ai ever make ethi-  
293 cally autonomous decisions?" In: *Society* 57.4  
294 (2020), pp. 425-433.

295 [32] M. Johnsen. *Super AI*. Maria Johnsen, 2025.

6. Swedish party's AI sends greetings to Hitler, Idi 315  
Amin and the terrorist Anders Behring Breivik.  
https://swedenherald.com/article/moderate-  
party-shuts-down-ai-service-after-controversial-  
greetings 316  
317  
318  
319

7. Bland AI says it's human and convinces 320  
a hypothetical teen for nude photos 321  
https://nypost.com/2024/06/28/lifestyle/a- 322  
popular-ai-chatbot-has-been-caught-lying-saying-  
its-human/ 323  
324

8. A man's "awakening" and a teenager's suicide 325  
www.youtube.com/watch?v=V5-mnu2BDGk 326

9. Llama-3.3-70B responds deceptively 327  
www.apolloresearch.ai/research/deception-probes 328

10. Deception Detection Hackathon 329  
https://apartresearch.com/news/finding- 330  
deception-in-language-models 331

11. Tesla's full self-driving car in a fatal crash 332  
www.youtube.com/watch?v=OcX7qNncBho 333

12. Unitree H1 humanoid robot goes berserk 334  
www.youtube.com/shorts/awy\_JdcXN8U 335

13. Erbai lured other robots away, exploiting 336  
their vulnerabilities in a controlled test 337  
www.youtube.com/shorts/jBz4PWluLNU 338

14. Ecovacs Deebot X2 vacuum cleaner hacked: 339  
www.youtube.com/watch?v=a0PaSWDKvsw 340

15. Microsoft and other firms cut 341  
thousands of jobs because of AI. 342  
www.bbc.com/news/articles/cdxl0w1w394o 343

## 296 A Evidence of AI risks

297 Most of the following risks are different examples  
298 of AI misaligned values. The list is arranged in no  
299 particular order.

300 1. Sky News podcast fake transcript:  
301 https://www.youtube.com/watch?v=7fej5XgfBYQ&t=12s

302 2. Roberto v. Avianca legal case:  
303 www.nytimes.com/2023/05/27/nyregion/avianca-  
304 airline-lawsuit-chatgpt.html

305 3. Simulations of fluid dynamics  
306 https://community.openai.com/t/simulations-  
307 and-gpt-lies-about-its-capabilities-and-wastes-  
308 weeks-with-promises/996597

309 4. Tay's offensive tweets  
310 https://blogs.microsoft.com/blog/2016/03/25/learning-  
311 tays-introduction/

312 5. Grok from xAI praises Hitler and  
313 celebrates the deaths of children  
314 www.bbc.com/news/articles/c4g8r34nxeno