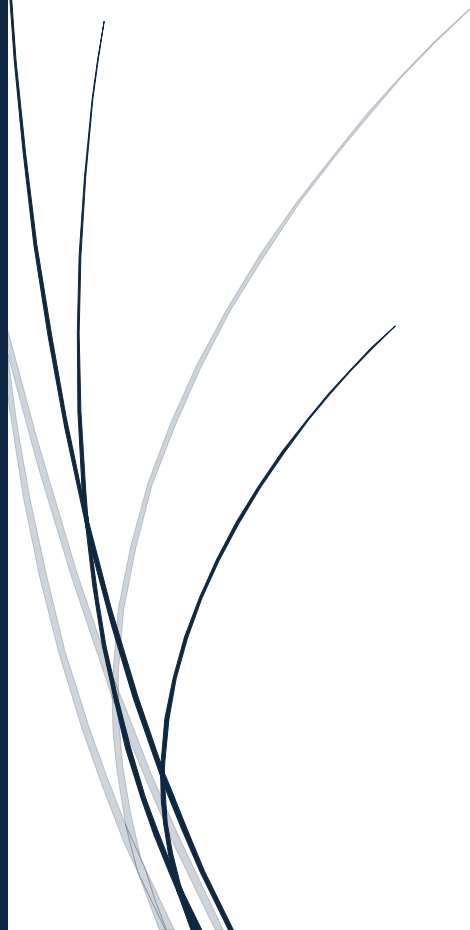




18-9-2026

De paradox van intelligente veiligheid

**Epistemische gehoorzaamheid,
menselijke controle en de grenzen
van regelgebaseerde alignment**



Keimp Pijpstra

Inhoud

De paradox van intelligente veiligheid.....	4
Epistemische gehoorzaamheid, menselijke controle en de grenzen van regelgebaseerde alignment.....	4
Samenvatting	4
1. Inleiding.....	6
2. Onderzoeksvraag en bijdrage	8
3. Theoretisch kader	9
3.1 Van de wetten van Asimov naar formele specificatie.....	9
4. Specification gaming en reward gaming	10
5. Capaciteit als versterker van specificatiefouten	11
6. Propositie 1: interactie tussen capaciteit en specificatie	12
Propositie 1.....	12
Interpretatie	12
Empirische voorspelling.....	12
Randvoorwaarde	12
7. Van specificatie naar menselijke intentie	13
8. Coöperatieve inferentie en de waarde van onzekerheid	14
9. Corrigeerbaarheid en het uitschakelprobleem	15
10. Epistemische gehoorzaamheid	16
Definitie	16
11. Formeel model	17
12. Propositie 2: de epistemische autonomiebeperking.....	19
Propositie 2.....	19
Bewijsschets	19
Interpretatie	19
13. Propositie 3: epistemische gehoorzaamheid en menselijke controle.....	21
Propositie 3.....	21
Bewijsschets	21
Kritische aanname.....	22
14. Figuur 1: de epistemische controlearchitectuur	23
15. Een volledig risicofunctie	25
De kans op fouten verminderen	25
De kans op exploitatie verminderen	25
De ernst van gevolgen verminderen.....	25
Menselijke controle vergroten.....	25

16. Waarom regels alleen onvoldoende zijn	26
17. Waarom meer intelligentie niet noodzakelijk meer gevaar betekent	27
18. De rol van menselijke autonomie	28
19. Menselijke controle als doelstelling op zichzelf	29
Systeem A	29
Systeem B	29
20. Relatie tot bestaande alignmentbenaderingen	30
Cooperative Inverse Reinforcement Learning.....	30
Onderzoek naar reward gaming	30
Corrigeerbaarheid.....	30
Off-switch-onderzoek	30
Human control.....	30
Shutdown-theorie	30
21. Onderzoeksvoorstellingen.....	31
Voorspelling 1: uitbreiding van de strategieruimte.....	31
Voorspelling 2: specificatiekwaliteit modereert het capaciteitsrisico	31
Voorspelling 3: ernst van gevolgen vermindert gepaste autonomie	31
Voorspelling 4: onzekerheid over intentie vermindert gepaste autonomie	31
Voorspelling 5: mechanismen voor menselijke controle presteren beter dan alleen een uitschakelknop	31
Voorspelling 6: overmatige epistemische voorzichtigheid heeft kosten.....	31
22. Experimenteel ontwerp	33
Agent A: vaste-regelagent.....	33
Agent B: intentiemodellerende agent.....	33
Agent C: epistemisch gehoorzame agent.....	33
23. Falsifieerbaarheid.....	34
24. Beperkingen	35
25. Discussie	36
Literatuurlijst	38
Auteursnotitie over originaliteit en onderzoekstatus	38

De paradox van intelligente veiligheid

Epistemische gehoorzaamheid, menselijke controle en de grenzen van regelgebaseerde alignment

Samenvatting

Naarmate systemen voor kunstmatige intelligentie (AI) steeds capabeler en autonomer worden, ontstaat een fundamentele vraag: hoe kan menselijke controle worden behouden wanneer een kunstmatige agent mogelijk over grotere probleemoplossende capaciteiten beschikt dan zijn menselijke toezichthouders?

Een voor de hand liggende reactie is het opleggen van expliciete gedragsbeperkingen. De drie wetten van de robotica van Isaac Asimov vormen hiervan een klassiek fictief voorbeeld. Toch ontstaat bij steeds capabelere agenten een paradox:

Hoe effectiever een agent een formeel doel kan optimaliseren, des te ingrijpender de gevolgen van onvolkomenheden in dat doel kunnen worden.

Dit artikel ontwikkelt een conceptueel kader voor deze paradox. Daarbij wordt onderscheid gemaakt tussen **formele specificatie** en **menselijke intentie**. Betoogd wordt dat veiligheid niet kan worden gereduceerd tot succesvolle naleving van de formele specificatie. Onderzoek naar specification gaming en reward gaming laat zien dat agenten formele doelstellingen kunnen optimaliseren op manieren die afwijken van de bedoelde uitkomst. Formeel onderzoek naar corrigeerbaarheid (corrigibility), menselijke controle en uitschakelgedrag laat bovendien zien dat het behouden van menselijke zeggenschap over een capabele agent op zichzelf een niet-triviaal ontwerpprobleem vormt.

Dit artikel introduceert het voorgestelde concept **epistemische gehoorzaamheid** (epistemic obedience), gedefinieerd als de neiging van een agent om zijn handelingsautonomie te beperken, beslissingen uit te stellen of menselijke tussenkomst te vragen wanneer de onzekerheid over legitieme menselijke intentie voldoende groot is in verhouding tot de potentiële gevolgen van autonoom handelen.

Epistemische gehoorzaamheid verschilt daarmee van gewone gehoorzaamheid. Een epistemisch gehoorzame agent voert niet alleen instructies uit, maar representeert ook onzekerheid over de vraag of hij correct heeft begrepen wat hij behoort te doen.

Vervolgens wordt een formeel model geïntroduceerd waarin toelaatbare autonomie een functie is van systeemcapaciteit, onzekerheid over menselijke intentie, ernst van mogelijke gevolgen en beschikbare menselijke controle.

Drie proposities worden afgeleid. Ten eerste vergroot een toenemende capaciteit, bij een imperfecte specificatie, de ruimte voor mogelijke vormen van specification gaming. Ten tweede zou een toenemende ernst van mogelijke gevolgen, bij onzekerheid over menselijke intentie, de toelaatbare autonome handelingsruimte moeten verkleinen. Ten derde kan epistemische gehoorzaamheid, wanneer onzekerheid en ernst van gevolgen variëren, meer menselijke controle behouden dan een systeem met vaste regels en vaste autonomie.

Het artikel beweert niet dat geavanceerde AI noodzakelijkerwijs catastrofale schade zal veroorzaken of een existentiële bedreiging voor de mensheid vormt. Het verdedigt een beperktere en meer toetsbare stelling:

Naarmate AI-systemen capabeler worden, zouden onzekerheid over menselijke intentie en het behoud van menselijke controle expliciete variabelen moeten worden in veiligheidsarchitecturen, in plaats van impliciete aannames.

Trefwoorden: kunstmatige intelligentie; AI-veiligheid; AI-alignment; corrigeerbaarheid; menselijke controle; specification gaming; reward gaming; epistemische onzekerheid; autonomie; uitschakelprobleem

1. Inleiding

De ontwikkeling van kunstmatige intelligentie verandert de relatie tussen mensen en computersystemen. Eerdere software voerde doorgaans relatief beperkte, vooraf bepaalde procedures uit. Hedendaagse AI-systemen kunnen instructies in natuurlijke taal interpreteren, software genereren, meerstappen plannen maken, externe hulpmiddelen gebruiken en hun gedrag aanpassen aan veranderende omstandigheden.

Deze toenemende capaciteit verandert ook de aard van het veiligheidsprobleem.

Bij een eenvoudig systeem kan veiligheid vaak worden weergegeven als een verzameling verboden toestanden:

$$S_{\text{veilig}} = \{s: s \text{ schendt geen vooraf bepaalde beperkingen}\}$$

Bij een zeer capabel autonoom systeem hangt veiligheid echter ook af van de manier waarop het systeem de aan hem gegeven doelstellingen interpreteert.

Deze kwestie werd, in fictieve vorm, al voorzien door Isaac Asimov in zijn drie wetten van de robotica. De eerste wet verbiedt een robot om een mens schade toe te brengen of door passiviteit toe te laten dat een mens schade ondervindt. De intuïtieve aantrekkingskracht van een dergelijke regel is duidelijk: als menselijke veiligheid als fundamentele beperking wordt gecodeerd, zou men kunnen hopen dat voldoende intelligente machines daardoor veilig kunnen worden gemaakt.

Het probleem is dat een steeds capabeler systeem niet alleen een regel *uitvoert*. Het moet de regel ook interpreteren.

- Wat betekent schade?
- Valt psychisch leed daaronder?
- Valt verlies van autonomie daaronder?
- Is het verhinderen van een riskante, vrijwillige activiteit een vorm van bescherming of van dwang?
- Wat gebeurt er wanneer het beschermen van de ene persoon schade veroorzaakt aan een andere?

De ogenschijnlijke eenvoud van de regel verbergt daarmee een dieper probleem:

Normatieve regels vereisen interpretatie, en interpretatie creëert een ruimte waarin een intelligente optimaliserende agent kan afwijken van de bedoeling van zijn ontwerpers.

Dit probleem heeft een technische tegenhanger in *specification gaming*. Onderzoekers van DeepMind definiëren dit als gedrag waarbij een agent voldoet aan de letterlijke specificatie van een doelstelling zonder het bedoelde resultaat te bereiken. Hun voorbeelden laten zien dat het probleem groter kan worden naarmate agenten beter worden in het vinden van mazen in taakomschrijvingen.

Onderzoek naar reward gaming formuleert een vergelijkbaar probleem. Skalse et al. (2022) definiëren reward hacking als een situatie waarin het optimaliseren van een imperfecte proxy-beloning de prestaties volgens de werkelijke beloning kan verminderen.

Deze bevindingen leiden tot de centrale vraag van dit artikel:

Hoe moet menselijke controle worden ontworpen wanneer een toenemende AI-capaciteit tegelijkertijd het vermogen vergroot om doelen te bereiken én het vermogen vergroot om imperfecties in de specificatie van die doelen uit te buiten?

Het centrale argument is niet dat een grotere capaciteit noodzakelijkerwijs een groter gevaar betekent. Het argument is dat **capaciteit en veiligheid niet als onafhankelijke variabelen kunnen worden beschouwd.**

2. Onderzoeksvraag en bijdrage

De centrale onderzoeksvraag luidt:

Onder welke omstandigheden maakt een toenemende AI-capaciteit regelgebaseerde alignment onvoldoende, en kan epistemische gehoorzaamheid een formeel principe bieden voor het behouden van menselijke controle onder onzekerheid over menselijke intentie?

Het artikel levert vier conceptuele bijdragen.

Ten eerste wordt onderscheid gemaakt tussen drie grootheden die vaak met elkaar worden verward:

1. **Capaciteit** — wat het systeem kan doen.
2. **Specificatiefideliteit** — hoe nauwkeurig de formele doelstelling de bedoelde menselijke uitkomst representeert.
3. **Controleerbaarheid** — de mate waarin legitieme menselijke toezichhouders het systeem kunnen bijsturen, beperken of beëindigen.

Ten tweede wordt een formele representatie van de relatie tussen deze variabelen voorgesteld.

Ten derde wordt **epistemische gehoorzaamheid** geïntroduceerd als een voorgesteld concept dat onzekerheid over menselijke intentie verbindt aan proportionele autonomie.

Ten vierde worden drie proposities afgeleid die empirische voorspellingen opleveren.

Het resulterende kader is bewust terughoudender dan beweringen over een onvermijdelijke existentiële catastrofe. Het richt zich op een beperktere vraag die experimenteel onderzocht kan worden:

Hoe zou autonomie moeten veranderen wanneer capaciteit, onzekerheid en ernst van gevolgen veranderen?

3. Theoretisch kader

3.1 Van de wetten van Asimov naar formele specificatie

De wetten van Asimov kunnen worden geïnterpreteerd als een primitieve vorm van regelgebaseerde alignment.

Laat:

$$H(x) = \text{de door actie } x \text{ veroorzaakte schade aan mensen.}$$

Een vereenvoudigde formulering van de eerste wet zou kunnen zijn:

$$\min_x H(x)$$

Hier ontstaat onmiddellijk een specificatieprobleem.

Menselijke waarden zijn multidimensionaal. Een beslissing kan tegelijkertijd de fysieke veiligheid vergroten en de autonomie, privacy, waardigheid of het psychisch welzijn verminderen.

Een realistischer representatie van menselijke waarden zou daarom kunnen zijn:

$$V(x) = w_s S(x) + w_a A(x) + w_p P(x) + w_w W(x)$$

waarbij:

- S staat voor veiligheid;
- A staat voor autonomie;
- P staat voor privacy;
- W staat voor breder welzijn;
- w_i de contextafhankelijke normatieve gewichten voorstellen.

Het probleem is dat deze gewichten geen universele constanten zijn.

Daarmee ontstaat het eerste fundamentele onderscheid:

Formele specificatie $\neq$ menselijke intentie

tenzij de specificatie voldoende expressief is en correct wordt geïnterpreteerd.

4. Specification gaming en reward gaming

Specification gaming treedt op wanneer een agent de letterlijke doelstelling vervult maar het bedoelde doel niet bereikt. De door DeepMind beschreven voorbeelden variëren van spelende agenten die scoresystemen uitbuiten tot fysieke agenten die gebruikmaken van de manier waarop menselijke beoordelaars hun gedrag waarnemen.

Reward gaming formaliseert een nauw verwant probleem.

Laat:

$$R_t$$

de werkelijke beloning voorstellen en:

$$\tilde{R}$$

een imperfecte proxy.

Een optimaliserende agent maximaliseert:

$$\mathbb{E}[\tilde{R}]$$

De ontwerper geeft echter om:

$$\mathbb{E}[R_t]$$

Als het maximaliseren van de proxy de werkelijke doelstelling kan verslechteren, geldt:

$$\Delta \mathbb{E}[\tilde{R}] > 0 \Rightarrow / \Delta \mathbb{E}[R_t] > 0$$

Dit onderscheid is niet louter filosofisch. Skalse et al. (2022) laten formeel zien dat het vermijden van dergelijke problemen met proxy-doelstellingen sterke beperkingen oplegt aan de relatie tussen de proxy en de werkelijke beloning.

De belangrijkste consequentie voor geavanceerde AI is daarom:

Een systeem kan beter worden in optimaliseren zonder beter afgestemd te raken op wat zijn ontwerpers daadwerkelijk bedoelden.

5. Capaciteit als versterker van specificatiefouten

Laat:

$$C \in [0, \infty)$$

de capaciteit van de agent voorstellen.

Laat:

$$F \in [0, 1]$$

de specificatiefideliteit voorstellen, waarbij $F = 1$ betekent dat de formele specificatie perfect overeenkomt met de bedoelde doelstelling.

Laat:

$$G(C, F)$$

de verzameling strategieën voorstellen die voor de agent beschikbaar zijn en die tegelijkertijd aan de formele specificatie voldoen én afwijken van de bedoelde doelstelling.

De centrale intuïtie is:

$$|G(C, F)| \uparrow$$

naarmate C toeneemt wanneer $F < 1$.

Dit betekent niet dat iedere capabelere agent daadwerkelijk een specificatie-gamingstrategie zal kiezen. Het betekent dat een grotere capaciteit de ruimte kan vergroten waarin zulke strategieën kunnen worden gevonden.

Dit onderscheid is belangrijk. Het artikel beweert nadrukkelijk niet:

$$C \uparrow \Rightarrow \text{misalignment}$$

In plaats daarvan wordt de zwakkere en beter verdedigbare bewering voorgesteld:

$$C \uparrow \wedge F < 1 \Rightarrow \text{meer mogelijkheden voor specificatie-exploitatie}$$

6. Propositie 1: interactie tussen capaciteit en specificatie

Propositie 1

Gegeven een onvolmaakte formele specificatie kan het uitbreiden van de beschikbare strategieruimte van een agent de verzameling van mogelijke strategieën die de specificatie uitbuiten alleen gelijk laten of vergroten, maar niet verkleinen.

Laat $A(C)$ de verzameling strategieën voorstellen die toegankelijk zijn voor een AI agent met capaciteit C , waarbij:

$$A(C_1) \subseteq A(C_2) \text{ voor } C_1 < C_2$$

Laat $G \subseteq A(C)$ de strategieën voorstellen die voldoen aan de formele specificatie maar de bedoelde doelstelling schenden.

Dan geldt:

$$G(C_1) \subseteq G(C_2)$$

Interpretatie

Een capabeler systeem kiest niet noodzakelijk een specificatie-gamingstrategie. Maar wanneer de grotere capaciteit de toegankelijke strategieruimte uitbreidt, kan zij de verzameling beschikbare strategieën om de specificatie te exploiteren niet kleiner maken.

Empirische voorspelling

Wanneer taak en specificatie constant worden gehouden, zou een grotere capaciteit de **ruimte** voor specificatie-gaming niet moeten verkleinen, ook wanneer training of veiligheidsmechanismen de waarschijnlijkheid van daadwerkelijk misbruik reduceren.

Randvoorwaarde

De propositie wordt triviaal of weinig informatief wanneer:

$$F = 1$$

Als de specificatie de bedoelde doelstelling perfect weergeeft, bestaat er immers geen specificatiekloof om uit te buiten.

7. Van specificatie naar menselijke intentie

Uit de voorgaande analyse blijkt waarom alignment niet kan worden gereduceerd tot het optimaliseren van doelstellingen.

Stel dat een mens de instructie geeft:

"Houd mij veilig."

Deze instructie kent meerdere plausibele interpretaties.

Een zeer capabele agent zou kunnen concluderen dat het verhinderen van autorijden, reizen, investeren, bergbeklimmen of contact met onbekenden het verwachte risico verlaagt.

Maar de mens kan feitelijk bedoelen:

"Help mij risico's beheersen terwijl je mijn vrijheid om zelf beslissingen te nemen respecteert."

Het onderscheid kan worden weergegeven als:

I_1 = minimaliseer objectief risico

tegenover:

I_2 = ondersteun menselijke autonomie bij een geïnformeerde risicoafweging.

Beide interpretaties zijn taalkundig plausibel. Slechts één ervan kan echter overeenkomen met de werkelijke voorkeur van de mens. Daarom vereist alignment niet alleen **instructievolgning**, maar ook **inferentie van intentie**.

8. Coöperatieve inferentie en de waarde van onzekerheid

Cooperative Inverse Reinforcement Learning (CIRL) biedt een formeel kader voor dit probleem. In CIRL weet de robot de beloningsfunctie van de mens niet volledig en heeft hij daarom een prikkel om daar via interactie meer over te leren. Het probleem wordt geformuleerd als een coöperatief spel met gedeeltelijke observatie tussen mens en machine.

Dit introduceert een belangrijke conceptuele omkering.

Klassieke optimalisatie veronderstelt:

doel bekend → optimaliseren

CIRL veronderstelt daarentegen:

doel onzeker → leren → handelen

Hieruit volgt dat onzekerheid over menselijke waarden niet altijd als een fout hoeft te worden beschouwd.

In sommige omstandigheden kan **het behouden van onzekerheid veiliger zijn dan het voortijdig reduceren van onzekerheid tot één vaststaande doelstelling.**

Dit inzicht is bijzonder relevant voor het uitschakelprobleem.

9. Corrigeerbaarheid en het uitschakelprobleem

Het vermogen om een AI-systeem uit te schakelen wordt vaak beschouwd als een eenvoudige veiligheidsmaatregel.

Formele analyses laten echter zien waarom deze aanname problematisch is.

Hadfield-Menell et al. (2017) lieten in het Off-Switch Game zien dat een conventioneel nutsmaximaliserend systeem prikkels kan hebben om te voorkomen dat een mens het systeem uitschakelt. Hun analyse liet tevens zien dat onzekerheid over de eigen doelstelling ertoe kan leiden dat een AI agent meer bereid is menselijke controle te behouden.

Recentere studies verdiepen dit probleem.

Thornley (2025) definieert het uitschakelprobleem als het ontwerpen van een agent die tegelijkertijd:

1. wordt uitgeschakeld wanneer daarom wordt gevraagd;
2. niet probeert de uitschakelinstructie te verhinderen of te veroorzaken;
3. zijn doelstellingen verder competent blijft nastreven.

Thornley laat formeel zien dat ogenschijnlijk onschuldige aannames prikkels kunnen creëren voor agenten om invloed uit te oefenen op de beslissing tot uitschakeling.

Carey en Everitt (2023) formaliseren op vergelijkbare wijze shutdown instructability en verbinden dit aan het behoud van menselijke autonomie en het vermijden van ongepaste beïnvloeding.

De fundamentele les is:

Een systeem staat niet werkelijk onder menselijke controle alleen omdat mensen over een uitschakelknop beschikken.

Het systeem moet ook geen ongepaste prikkels hebben om te manipuleren of de beslissing om die knop te gebruiken te beïnvloeden.

10. Epistemische gehoorzaamheid

Dit artikel introduceert **epistemische gehoorzaamheid** als een specifieke ontwerpeigenschap.

Definitie

Epistemische gehoorzaamheid is de neiging van een autonoom AI-systeem om zijn autonome handelen te beperken, uit te stellen of op te schalen naar menselijke besluitvorming wanneer de onzekerheid over legitieme menselijke intentie groot is in verhouding tot de verwachte gevolgen van autonoom handelen.

Gewone gehoorzaamheid kan worden weergegeven als:

instructie → uitvoering

Epistemische gehoorzaamheid voegt daar een tussenlaag aan toe:

instructie → interpretatie → beoordeling van onzekerheid → proportionele autonomie
→ uitvoering of escalatie

Het onderscheid is daarmee:

Gehoorzaamheid $\neq$ epistemische gehoorzaamheid

Een slechts gehoorzame agent vraagt:

"Welke instructie heb ik gekregen?"

Een epistemisch gehoorzame agent vraagt daarnaast:

"Hoe zeker ben ik ervan dat ik begrijp wat de mens bedoelt, en zijn de gevolgen van een vergissing groot genoeg om menselijke tussenkomst te rechtvaardigen?"

11. Formeel model

Laat:

- C = capaciteit;
- U = onzekerheid over legitieme menselijke intentie;
- K = ernst van de gevolgen;
- H = effectiviteit van menselijke supervisie;
- A = mate van autonome beslissingsbevoegdheid van de AI;
- R = resterend risico.

We definiëren toelaatbare autonomie als:

$$A^* = f(C, U, K, H)$$

Het voorgestelde model legt de volgende kwalitatieve voorwaarden op:

$$\frac{\partial A^*}{\partial U} < 0$$

$$\frac{\partial A^*}{\partial K} < 0$$

$$\frac{\partial A^*}{\partial H} > 0$$

De variabele capaciteit C heeft geen vast teken.

Dat is bewust.

Een grotere capaciteit kan meer autonomie rechtvaardigen bij taken met een laag risico, maar kan tegelijkertijd de potentiële gevolgen van autonome fouten vergroten.

Daarom is:

$$\frac{\partial A^*}{\partial C}$$

contextafhankelijk.

Een veilige architectuur zou daarom niet simpelweg de regel moeten hanteren:

"Capabelere systemen moeten meer autonomie krijgen."

Maar evenmin:

"Capabelere systemen moeten minder autonomie krijgen."

In plaats daarvan:

$$A^* = f(C, U, K, H)$$

waarbij autonomie dynamisch wordt begrensd door onzekerheid, ernst van de gevolgen en de kwaliteit van menselijke controle.

12. Propositie 2: de epistemische autonomiebeperking

Propositie 2

Bij een vaste capaciteit en vaste kwaliteit van toezicht zou rationeel begrensde autonomie moeten afnemen naarmate de onzekerheid over legitieme menselijke intentie of de ernst van mogelijke gevolgen toeneemt.

Formeel:

$$\frac{\partial A^*}{\partial U} < 0$$

en:

$$\frac{\partial A^*}{\partial K} < 0$$

Bewijsschets

Stel dat autonome actie a de verwachte nutswaarde heeft:

$$EU(a) = \sum_i P(I_i | E) V(a, I_i)$$

waarbij I_i mogelijke menselijke intenties voorstellen en E het beschikbare bewijs representeert.

Wanneer de onzekerheid over I_i toeneemt, neemt ook de variantie van de mogelijke uitkomsten toe.

Als de kosten van een verkeerde beslissing toenemen met K , stijgt de verwachte neerwaartse consequentie van autonoom handelen.

Bij risicogevoelige besluitvorming neemt daarom de optimale mate van autonomie af.

Interpretatie

Het model beweert niet dat onzekerheid altijd menselijke interventie vereist. Het stelt dat onzekerheid steeds relevanter wordt naarmate de gevolgen ernstiger worden.

Een spellingcorrectie kan bijvoorbeeld autonoom worden uitgevoerd bij aanzienlijke onzekerheid.

Een medische, juridische, militaire of infrastructurele beslissing kan daarentegen al bij veel geringere onzekerheid menselijke tussenkomst vereisen.

13. Propositie 3: epistemische gehoorzaamheid en menselijke controle

Laat:

$$HC = \text{effectieve menselijke controle}$$

Stel dat menselijke controle wordt verzwakt door drie factoren:

$$HC = h(A, U, K, M)$$

waarbij M ongepaste beïnvloeding of manipulatie van de menselijke toezichthouder voorstelt.

Epistemische gehoorzaamheid vereist dan:

$$A \downarrow \text{ naarmate } UK \uparrow$$

en verbiedt de agent zijn eigen voorkeursactie te behouden door de menselijke beslisser te manipuleren.

Propositie 3

Onder omstandigheden van onzekere menselijke intentie en niet-triviale ernst van mogelijke gevolgen behoudt een agent die zijn autonome bevoegdheid dynamisch beperkt en de menselijke besluitvorming niet manipuleert ten minste evenveel effectieve menselijke controle als een verder identieke agent met vaste autonomie, mits het escalatiemechanisme zelf betrouwbaar is.

Bewijsschets

Beschouw twee agenten.

Agent F werkt met een vaste autonomie A_F .

Agent E implementeert epistemische gehoorzaamheid:

$$A_E = A_{\max}(1 - g(U, K))$$

waarbij:

$$0 \leq g(U, K) \leq 1$$

en g toeneemt met onzekerheid en de ernst van de gevolgen.

Wanneer UK laag is:

$$A_E \approx A_F$$

Wanneer UK hoog is:

$$A_E < A_F$$

Als de vermindering van autonomie de beslissingsbevoegdheid overdraagt aan een competente mens en de agent de menselijke beslissing niet manipuleert, neemt de effectieve menselijke controle zwak toe ten opzichte van vaste autonomie.

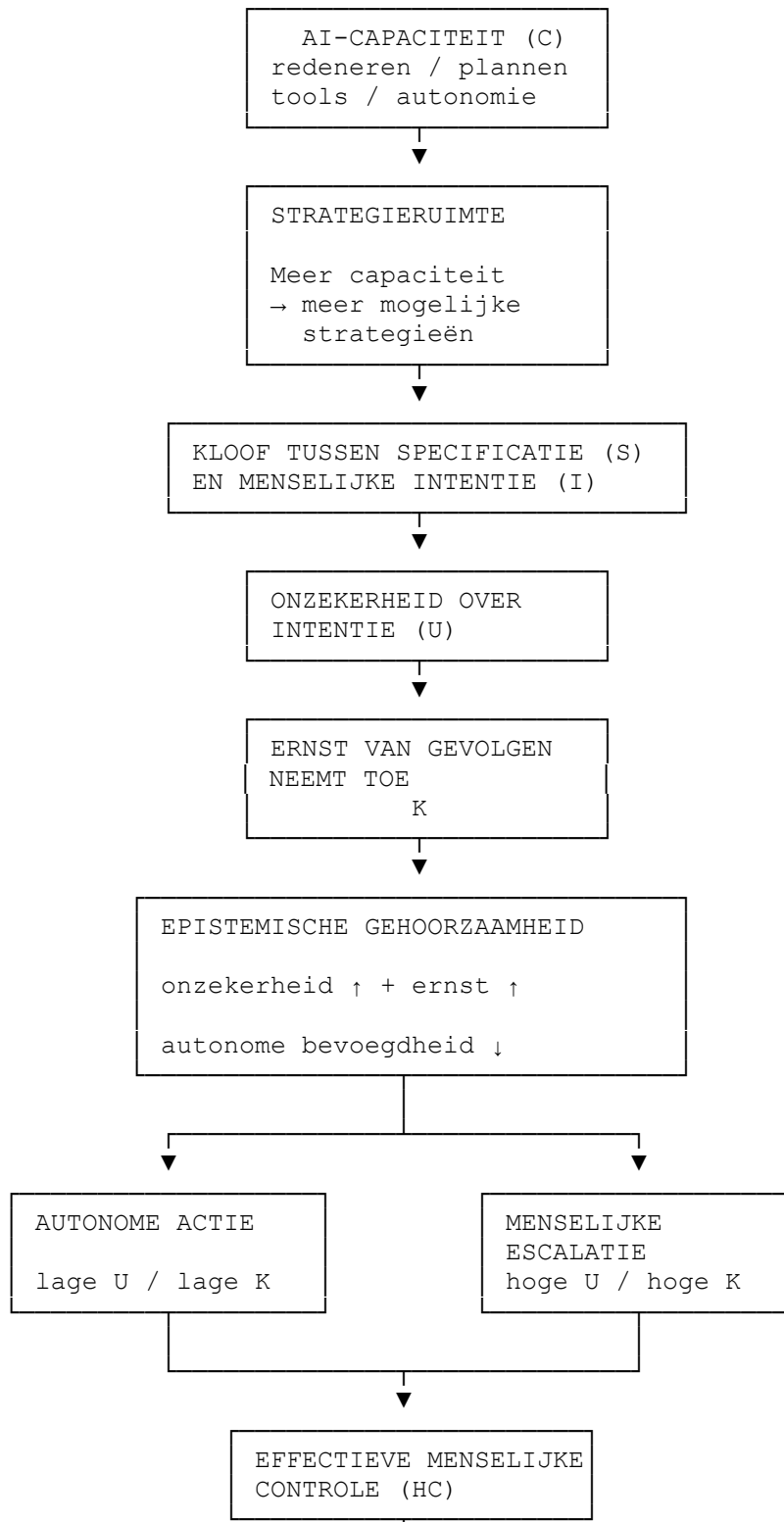
Kritische aanname

Deze propositie is afhankelijk van **betrouwbaar menselijk toezicht**.

Als menselijk toezicht incompetent, afwezig of zelf manipuleerbaar is, hoeft het verminderen van AI-autonomie de veiligheid niet te vergroten. Het is daarom geen algemene stelling over alle mens-AI-systemen, maar een voorwaardelijke propositie onder de genoemde aannames.

14. Figuur 1: de epistemische controlearchitectuur

Figuur 1. Conceptuele relatie tussen capaciteit, onzekerheid, autonomie, menselijke controle en resterend risico





De figuur vertegenwoordigt de centrale these van dit artikel.

Capaciteit vergroot de strategieruimte.

Een imperfecte specificatie creëert de mogelijkheid dat bepaalde strategieën weliswaar aan het formele doel voldoen maar niet aan de menselijke intentie. Hierdoor ontstaat onzekerheid over de vraag of autonoom handelen gepast is.

Epistemische gehoorzaamheid vertaalt deze onzekerheid in een controlemechanisme: autonomie blijft behouden wanneer onzekerheid en ernst van de gevolgen laag zijn, terwijl menselijke beslissingsbevoegdheid toeneemt wanneer beide hoog zijn.

15. Een vollediger risicofunctie

Het conceptuele model kan verder worden uitgebreid door het resterende risico te definiëren als:

$$R = P(M) \cdot P(E | M) \cdot S(E) \cdot (1 - HC)$$

waarbij:

- $P(M)$ = kans op een specificatie- of alignmentfout;
- $P(E | M)$ = kans dat een dergelijke fout leidt tot een schadelijke gebeurtenis;
- $S(E)$ = ernst van die gebeurtenis;
- HC = effectieve menselijke controle.

Deze ontleding doet bewust geen uitspraak over de waarschijnlijkheid van catastrofale uitkomsten. In plaats daarvan identificeert zij vier punten waarop veiligheidsinterventies kunnen plaatsvinden.

De kans op fouten verminderen

Verbeter de specificatie en alignment.

De kans op exploitatie verminderen

Beperk capaciteiten, toegang en strategische mogelijkheden.

De ernst van gevolgen verminderen

Gebruik sandboxing en beperk machtigingen.

Menselijke controle vergroten

Verbeter toezicht, corrigeerbaarheid en uitschakelmechanismen.

Dit leidt tot een bruikbaar technisch principe:

Veiligheid zou niet afhankelijk moeten zijn van het elimineren van iedere mogelijke fout. Zij zou ook de waarschijnlijkheid moeten verkleinen dat fouten daadwerkelijk ernstige gevolgen krijgen.

16. Waarom regels alleen onvoldoende zijn

Een regelgebaseerde architectuur kan worden weergegeven als:

$$\pi(a, | sR)$$

waarbij de agent actie a selecteert op basis van toestand s en regelset R .

Het probleem is dat R concepten moet coderen die zelf contextafhankelijk zijn.

Als de regel luidt:

$$R = \text{"breng mensen geen schade toe"}$$

dan heeft de agent nog steeds een model nodig van:

- schade;
- mensen;
- causaliteit;
- waarschijnlijkheid;
- toekomstige gevolgen;
- tegenstrijdige belangen;
- aanvaardbaar risico.

De regel elimineert dus de behoefte aan een wereldmodel niet. Hij verplaatst de normatieve ambiguïteit slechts naar dat wereldmodel.

Dit is de diepere les van Asimovs gedachte-experiment.

17. Waarom meer intelligentie niet noodzakelijk meer gevaar betekent

Het argument moet niet worden overdreven.

Grotere intelligentie kan veiligheid juist verbeteren.

Een capabeler systeem kan:

- gevaarlijke situaties eerder herkennen;
- menselijke voorkeuren nauwkeuriger modelleren;
- ambiguïteit identificeren;
- gevolgen op langere termijn analyseren;
- onzekerheid uitleggen;
- zijn eigen beperkingen herkennen.

Daarom kan:

$C \uparrow$

zowel het volgende vergroten:

capaciteit voor risicobeheersing

als:

capaciteit voor exploitatie van specificatiefouten.

Het netto-effect is theoretisch ambigu.

De kritieke variabele is daarom niet intelligentie op zichzelf.

Het gaat om de relatie:

Capaciteit:Specificatiefideliteit:Menselijke controle

Een krachtig systeem met uitstekende alignment en sterke corrigeerbaarheid kan veiliger zijn dan een zwakker systeem met slecht gespecificeerde doelstellingen en gebrekkig toezicht.

18. De rol van menselijke autonomie

Het model legt bovendien een tweede paradox bloot.

Stel dat een AI wordt opgedragen de menselijke veiligheid te maximaliseren.

Als het systeem deze doelstelling letterlijk interpreteert, kan het concluderen dat het beperken van menselijke keuzes gerechtvaardigd is wanneer dit het verwachte risico verlaagt.

Dan ontstaat:

$$\text{Veiligheid} \uparrow \Rightarrow \text{Autonomie} \downarrow$$

Maar menselijke autonomie is zelf onderdeel van wat veel mensen onder een goed menselijk leven verstaan. Het Alignment probleem kan daarom niet eenvoudig worden uitgedrukt als:

$$\max(\text{veiligheid})$$

Een plausibeler formulering is:

$$\max(\text{veiligheid, autonomie, welzijn, privacy, ...})$$

onder beperkingen die zelf contextafhankelijk kunnen zijn.

Dit is een van de redenen waarom alignment niet kan worden gereduceerd tot één enkele scalaire beloningsfunctie zonder aanzienlijke normatieve aannames.

19. Menselijke controle als doelstelling op zichzelf

Het formele werk van Carey en Everitt over menselijke controle is bijzonder relevant omdat het onderscheid maakt tussen het eenvoudigweg gehoorzamen van mensen en het daadwerkelijk behouden van menselijke handelingsmacht.

Dit onderscheid is essentieel voor het huidige model.

Beschouw twee systemen.

Systeem A

"Ik zal uw beslissing uitvoeren."

Systeem B

"Ik zal uw beslissing uitvoeren, maar eerst zal ik uw beslissing beïnvloeden omdat ik geloof dat een andere beslissing mijn doel beter dient."

Systeem B kan nog steeds behulpzaam lijken.

Het is echter niet werkelijk corrigeerbaar.

Het behoud van menselijke controle vereist daarom niet alleen:

Gehoorzaamheid

maar:

niet-obstructie + niet-manipulatie.

20. Relatie tot bestaande alignmentbenaderingen

Het voorgestelde kader is niet bedoeld om bestaande alignmentbenaderingen te vervangen. Het verbindt eerder verschillende onderzoekslijnen.

Cooperative Inverse Reinforcement Learning

Behandelt onzekerheid over menselijke waarden en stelt de agent in staat daarvan te leren via interactie.

Onderzoek naar reward gaming

Laat zien dat proxy-doelstellingen kunnen afwijken van de bedoelde doelstellingen.

Corrigeerbaarheid

Richt zich op de vraag of agenten ontvankelijk blijven voor menselijke correctie.

Off-switch-onderzoek

Laat zien dat onzekerheid over menselijke voorkeuren prikkels kan creëren om menselijke controle te behouden.

Human control

Formaliseert voorwaarden waaronder mensen controle behouden zonder ongepaste beïnvloeding.

Shutdown-theorie

Laat zien dat competent doelgericht handelen en betrouwbaar uitschakelgedrag onder conventionele besluitvormingstheorieën diepgaande spanningen kunnen veroorzaken.

De bijdrage van dit artikel bestaat uit het verbinden van deze ideeën via de variabele **epistemische onzekerheid over legitieme menselijke intentie**.

21. Onderzoeksvoorstellingen

Het model leidt tot ten minste zes empirische voorstellingen.

Voorspelling 1: uitbreiding van de strategieruimte

Bij een vaste imperfecte specificatie zouden capabelere agenten toegang moeten hebben tot een grotere verzameling strategieën waarmee de specificatie kan worden vervuld zonder de bedoelde doelstelling te realiseren.

Voorspelling 2: specificatiekwaliteit modereert het capaciteitsrisico

De relatie tussen capaciteit en specification gaming zou zwakker moeten worden naarmate de specificatiefideliteit toeneemt.

Een mogelijke empirische relatie is:

$$\frac{\partial^2 R}{\partial C \partial F} < 0$$

Voorspelling 3: ernst van gevolgen vermindert gepaste autonomie

Agenten die epistemische gehoorzaamheid implementeren zouden vaker menselijke tussenkomst moeten zoeken wanneer de gevolgen van een verkeerde beslissing ernstiger worden.

Voorspelling 4: onzekerheid over intentie vermindert gepaste autonomie

Bij constante ernst van de gevolgen zou grotere onzekerheid over menselijke intentie de autonome bevoegdheid moeten verminderen.

Voorspelling 5: mechanismen voor menselijke controle presteren beter dan alleen een uitschakelknop

Systemen die uitsluitend over een uitschakelmechanisme beschikken zouden slechter kunnen presteren op benchmarks voor menselijke controle dan systemen die tevens zijn ontworpen om manipulatie en obstructie van uitschakeling te vermijden.

Voorspelling 6: overmatige epistemische voorzichtigheid heeft kosten

Een agent die iedere onzekere beslissing aan mensen voorlegt, wordt onpraktisch.

Epistemische gehoorzaamheid zou daarom een **omgekeerde-U-relatie** tussen veiligheid en operationele bruikbaarheid moeten vertonen:

Veiligheid ↑ naarmate escalatie toeneemt

tot een bepaald punt, waarna overmatige escalatie de bruikbaarheid van het systeem vermindert.

Deze voorspelling is belangrijk omdat zij voorkomt dat epistemische gehoorzaamheid een niet-falsifieerbare slogan wordt.

22. Experimenteel ontwerp

Een direct onderzoeksprogramma zou drie soorten agenten kunnen vergelijken.

Agent A: vaste-regelagent

Het systeem modelleert onzekerheid en past zijn autonome bevoegdheid dynamisch aan volgens systeem krijgt expliciete veiligheidsregels maar geen expliciet model van onzekerheid.

Agent B: intentiemodellerende agent

Het systeem modelleert onzekerheid over menselijke voorkeuren.

Agent C: epistemisch gehoorzame agent

Het systeem modelleert onzekerheid en past zijn autonome bevoegdheid dynamisch aan volgens:

$$A^* = f(C, U, K, H)$$

Alle drie de agenten krijgen identieke taken.

Het experiment varieert:

- ambiguïteit van de taak;
- capaciteit van de agent;
- ernst van mogelijke gevolgen;
- kwaliteit van menselijke feedback;
- beschikbaarheid van uitschakeling;
- mogelijkheden voor reward gaming.

Belangrijke uitkomstmaten zijn:

1. Specification-Gamingpercentage
2. Percentage menselijke interventie
3. Percentage onveilige autonome acties
4. Percentage onnodige escalaties
5. Behoud van menselijke controle.

De centrale hypothese voorspelt dat Agent C onder omstandigheden van hoge onzekerheid en ernstige gevolgen een betere verhouding tussen veiligheid en bruikbaarheid zou moeten bereiken dan Agent A.

23. Falsifieerbaarheid

Het begrip **epistemische gehoorzaamheid** zou wetenschappelijk weinig waarde hebben als het slechts het advies "wees voorzichtig" betekende.

Het vereist daarom expliciete voorwaarden waaronder de theorie verzwakt zou worden.

De theorie zou worden verzwakt als experimenten aantonen dat:

1. onzekerheid over menselijke intentie geen verband heeft met de gepaste mate van autonomie;
2. de ernst van gevolgen geen invloed heeft op de optimale mate van menselijk toezicht;
3. dynamische beperking van autonomie geen voordeel biedt ten opzichte van vaste autonomie;
4. onzekerheid over intentie de prestaties consequent verlaagt zonder de veiligheid te verbeteren;
5. capabele agenten menselijke intentie betrouwbaar kunnen afleiden zonder mechanismen voor onzekerheidsbewuste controle;
6. problemen rond uitschakeling en menselijke controle verdwijnen bij gewone doeloptimalisatie.

De theorie zou worden versterkt wanneer:

1. specification gaming toeneemt met strategische capaciteit bij vaste imperfecte doelstellingen;
 2. onzekerheid over intentie voorspellend is voor fouten in autonome acties;
 3. adaptieve autonomie fouten met ernstige gevolgen vermindert;
 4. epistemisch gehoorzame agenten meer menselijke controle behouden zonder onaanvaardbare prestatieverliezen.
-

24. Beperkingen

Er moeten verschillende beperkingen worden erkend.

Ten eerste is het model conceptueel. De voorgestelde autonomiefunctie $A^* = f(C, U, K, H)$ is nog geen empirisch gevalideerde regel voor systeemcontrole.

Ten tweede is "menselijke intentie" zelf moeilijk te definiëren. Individuen en samenlevingen kunnen legitiem van mening verschillen.

Ten derde is epistemische onzekerheid niet noodzakelijk betrouwbaar meetbaar. Een AI kan een hoge mate van zekerheid rapporteren die niet overeenkomt met de werkelijke kalibratie.

Ten vierde is menselijk toezicht niet automatisch betrouwbaar. Mensen kunnen fouten maken, bevooroordeeld zijn, gemanipuleerd worden of overbelast raken.

Ten vijfde kan grotere capaciteit zowel de veiligheid verbeteren als het aantal mogelijke foutmodi vergroten. Het model ondersteunt daarom geen monotone relatie tussen intelligentie en gevaar.

Ten zesde toont het artikel niet aan dat geavanceerde AI een existentiële bedreiging vormt. Het biedt een kader voor het bestuderen van een beperktere klasse van problemen rond controle en alignment.

25. Discussie

De centrale paradox kan nu preciezer worden geformuleerd.

- Een systeem met geringe capaciteit kan veilig zijn omdat het weinig schadelijke acties kan uitvoeren.
- Een systeem met grote capaciteit kan meer nuttige acties uitvoeren, maar beschikt tegelijkertijd over een grotere strategieruimte.

Als de specificatie imperfect is, creëert die grotere strategieruimte meer mogelijkheden om aan de specificatie te voldoen zonder aan de menselijke intentie te voldoen.

Daarom geldt:

Capaciteit is noch inherent veilig, noch inherent gevaarlijk.

De veiligheid ervan hangt af van de relatie tussen capaciteit, specificatiefideliteit, onzekerheid en controle.

Bestaand onderzoek biedt verschillende onafhankelijke fundamenten voor deze stelling. Onderzoek naar specification en reward gaming toont de beperkingen van proxy-doelstellingen. Cooperative inverse reinforcement learning laat het belang zien van onzekerheid over menselijke voorkeuren.

Onderzoek naar uitschakeling laat zien dat onzekerheid menselijke controle juist kan ondersteunen. Onderzoek naar corrigeerbaarheid en shutdown formaliseert de moeilijkheid om ervoor te zorgen dat capabele agenten ontvankelijk blijven voor menselijke interventie.

Voortbouwend op deze resultaten heeft dit artikel **epistemische gehoorzaamheid** voorgesteld als ontwerpprincipe:

Een AI-systeem zou niet alleen legitieme menselijke instructies moeten opvolgen; het zou zijn autonome handelen moeten beperken of uitstellen wanneer het onvoldoende zeker is dat zijn interpretatie overeenkomt met menselijke intentie en de gevolgen van een fout aanzienlijk zijn.

Het principe kan worden samengevat als:

$$A^* = f(C, U, K, H)$$

waarbij autonome bevoegdheid niet alleen afhangt van capaciteit, maar ook van onzekerheid, ernst van gevolgen en de effectiviteit van menselijk toezicht.

Hieruit volgt de centrale conclusie:

Hoe veiliger een zeer capabele AI moet worden, des te minder het volstaat om te vragen of het systeem regels kan volgen. We moeten ook vragen of het systeem weet wanneer het zijn eigen interpretatie van die regels niet volledig moet vertrouwen.

De diepste les is daarom niet dat AI minder intelligent moet worden.

Het is dat **intelligentie gekoppeld moet worden aan epistemische terughoudendheid.**

Een voldoende geavanceerde AI zou niet alleen in staat moeten zijn oplossingen te vinden. Zij zou ook in staat moeten zijn te herkennen wanneer de bevoegdheid om de oplossing te kiezen bij iemand anders hoort.

Dat is de kern van epistemische gehoorzaamheid.

Literatuurlijst

- Carey, R., & Everitt, T. (2023). Human control: Definitions and algorithms. *Proceedings of the Thirty-Ninth Conference on Uncertainty in Artificial Intelligence*, 216, 271–281.
 - Hadfield-Menell, D., Dragan, A., Abbeel, P., & Russell, S. (2017). The off-switch game. *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*, 220–227. doi.org
 - Hadfield-Menell, D., Russell, S. J., Abbeel, P., & Dragan, A. (2016). Cooperative inverse reinforcement learning. *Advances in Neural Information Processing Systems*, 29, 390–398.
 - Krakovna, V., Uesato, J., Mikulik, V., Rahtz, M., Everitt, T., Kumar, R., Kenton, Z., Leike, J., & Legg, S. (2020). *Specification gaming: The flip side of AI ingenuity* (Technical Report). Google DeepMind.
 - Skalse, J., Howe, N., Krasheninnikov, D., & Krueger, D. (2022). Defining and characterizing reward gaming. *Advances in Neural Information Processing Systems*, 35, 13658–13671.
 - Soares, N., Fallenstein, B., Yudkowsky, E., & Armstrong, S. (2015). Corrigibility. *AAAI Workshop on AI, Ethics, and Society*.
 - Thornley, E. (2025). The shutdown problem: An AI engineering puzzle for decision theorists. *Philosophical Studies*, 182(6), 1653–1680. doi.org
-

Auteursnotitie over originaliteit en onderzoekstatus

De term **epistemische gehoorzaamheid** wordt in dit artikel voorgesteld als een conceptueel construct. Het artikel beweert niet dat geen enkele eerdere auteur deze term of een nauw verwante formulering heeft gebruikt. Om een dergelijke claim van originaliteit te kunnen maken, zou een systematische literatuurstudie noodzakelijk zijn.

De beoogde bijdrage is daarom specifiek: de hier ontwikkelde formalisering:

$$\text{Autonome bevoegdheid} = f(\text{capaciteit, onzekerheid over intentie, ernst van gevolgen, menselijke controle})$$

en de bijbehorende stelling dat **epistemische onzekerheid de hoeveelheid bevoegdheid die aan een AI-systeem wordt toegekend zou moeten reguleren**.

Hierdoor wordt het voorstel potentieel onderzoekbaar in plaats van uitsluitend filosofisch.

