Theoretisch kader

Onafhankelijke toetsing *Moral Knight* februari 2026

Hieronder volgt de uitwerking van belangrijke argumenten en het visiestuk (theoretisch kader), voorzien van de correcte APA-verwijzingen en een literatuurlijst gebaseerd op relevante en actuele academische bronnen.

Deel 1:
Selectie van zwaarwegende argumenten voor onafhankelijke toetsing

---

Op basis van de bronnen zijn dit de vijf meest zwaarwegende argumenten voor onafhankelijke toetsing van AI in het publieke domein.

1. **Tegengaan van het 'Ongereguleerde wilde westen':**

Het uitrollen van AI-technologieën in de echte wereld ('in-the-wild') vindt momenteel vaak plaats zonder adequaat toezicht, duidelijke governance of ethische waarborgen. Onafhankelijke toetsing is noodzakelijk om juridische, ethische en epistemische verantwoordelijkheid af te dwingen en te voorkomen dat burgers onderwerp worden van ongereguleerde experimenten die hun rechten kunnen schenden. Hierbij is het cruciale verschil tussen verantwoord testen en de tegenwoordige realiteit van het 'wilde westen' dat live tests idealiter eerst plaatsvinden binnen gecontroleerde kaders (sandboxes) met voorafgaand toezicht. Niet achteraf tijdens een ongereguleerde uitrol op burgers (Yeung & Li, 2025). Vergelijkbaar met de farmacie hebben we een 'fase IV'-benadering nodig. Een ‘post-deployment monitoring’ waarbij we niet stoppen met toetsen na de lab-fase, maar de impact in de echte wereld blijven monitoren (Spiegelhalter, 2020, zoals geciteerd in Yeung & Li, 2025).

2. **Overbruggen van de 'Evaluation gap' en de kloof tussen theorie en praktijk:**

Er is een aantoonbare kloof tussen hoe AI presteert in gecontroleerde testomgevingen (benchmarks) en hoe het functioneert in de complexe werkelijkheid. Dit wordt de ‘evaluation gap’ genoemd (International AI Safety Report, 2026). Zonder onafhankelijke toetsing in de praktijkcontext blijven onvoorziene risico's en faalmodi onopgemerkt, omdat abstracte ethische

principes zich niet automatisch vertalen naar veilige toepassingen (Herzog & Blank, 2024).

3. **Waarborgen van operationele effectiviteit (Live output):**

Een systeem kan technisch goed functioneren ('functional performance'), maar in de praktijk falen om het beoogde maatschappelijke doel te bereiken ('operational effectiveness'). Toetsing moet zich daarom ook altijd richten op de *live output* waarmee de burger wordt geconfronteerd, omdat alleen daar zichtbaar wordt of een eventuel defect of inbreuk op rechten proportioneel is ten opzichte van het daadwerkelijk behaalde nut (Yeung & Li, 2025). Dit wordt extra urgent bij Generatieve AI (*GPAI*). Omdat de 'black box' die ten grondslag ligt aan een publieke AI-toepassing weinig opties overlaat om grip te krijgen op de onvoorspelbaarheid en ondoorzichtigheid van de techniek (International AI Safety Report, 2026).

4. **Bescherming van mensenrechten en fundamentele vrijheden:**

De inzet van algoritmes door de overheid raakt direct aan fundamentele rechten zoals non-discriminatie, privacy en behoorlijk bestuur. Instrumenten zoals het 'Impact Assessment Mensenrechten en Algoritmes' (*IAMA*) zijn essentieel om deze rechten systematisch te beschermen en te voorkomen dat technologieën bestaande ongelijkheid versterken (Gerards et al., 2026).

5. **Tegengaan van instrumentalisme en 'Ethics washing':**

Interne ethische reflectie van organisaties riskeert te verworden tot een afvink-exercitie ('box-ticking') om uitrol van AI te versnellen. Zonder dat de wenselijkheid van de technologie fundamenteel wordt bevraagd. Externe toetsing en 'applied ethics' zijn nodig om blinde vlekken te identificeren en machtsstructuren te bevragen, in plaats van technologie als enige oplossing te zien (Bleher & Braun, 2023; Herzog & Blank, 2024).

Deel 2:
Visiestuk (theoretisch kader)

Van principe naar praktijk: Een theoretisch kader voor de onafhankelijke toetsing van publieke AI

De digitalisering van de overheid brengt een fundamentele verschuiving teweeg in de relatie tussen de staat en de burger. Waar besluitvorming wordt gedelegeerd aan algoritmes en AI-systemen, volstaat het niet langer om te vertrouwen op goede intenties of abstracte principes. Om het vertrouwen in de digitale overheid te waarborgen, is een robuust stelsel van onafhankelijke toetsing noodzakelijk. Dit kader schetst de urgentie en de voorwaarden voor dergelijke toetsing.

Tot slot, bij het overbruggen van de kloof tussen theorie en praktijk ten aanzien van AI-ethiek is het nuttig het volgende onderscheid te maken. Namelijk, tussen 'ethiek in AI' - de technische code -, 'ethiek van AI' - de sociale praktijk - en 'ethiek én AI' - de principiële (on-) wenselijkheid - (Ratti, 2023). Een valide en betrouwbare onafhankelijke toets van publieke AI-toepassingen dient deze drie dimensies in overweging te nemen.

Het einde van het 'Ongereguleerde Wilde westen'

De huidige praktijk van AI-ontwikkeling en -implementatie in de publieke ruimte vertoont kenmerken van een ‘ongereguleerd Wilde Westen’ (Yeung & Li, 2025). Technologieën worden regelmatig 'in het wild' getest op burgers zonder dat daar adequate juridische of ethische kaders aan ten grondslag liggen. Dit brengt risico's met zich mee, variërend van discriminatie tot onrechtvaardige inbreuken op privacy. Onafhankelijke toetsing moet fungeren als een *guardrail* die garandeert dat experimenten en toepassingen niet alleen technisch functioneren, maar ook juridisch, ethisch en epistemisch verantwoord zijn (OECD, 2024).

De kloof tussen theorie en praktijk overbruggen

Er bestaat inmiddels een wereldwijde ‘consensus’ op het gebied van de waarden die relevant zijn ten aanzien van AI. Richtinggevende principes zoals *transparantie, rechtvaardigheid en eerlijkheid, niet-schaden, verantwoordelijkheid en privacy* zijn algemeen aangenomen (Jobin et al., 2019, zoals geciteerd in Dignum, 2022). Echter, de vertaling van abstracte principes naar de praktijk blijkt problematisch. Over de interpretatie en

praktische invulling van waarden en de weging hiervan bestaat allerminst overeenstemming (Dignum, 2022). Dit fenomeen, bekend als de *principles-to-practice gap*, zorgt ervoor dat ethische richtlijnen vaak weinig bescherming bieden tegen daadwerkelijke schade (Herzog & Blank, 2024). Recente inzichten tonen aan dat benchmarks in laboratoria vaak niet voorspellen hoe een systeem zich in de echte wereld gedraagt, wat leidt tot een ‘evaluation gap’ (International AI Safety Report, 2026).

Toegepaste ethiek als proces

Om deze kloof te dichten, moet onafhankelijke toetsing fungeren als een vorm van *toegepaste ethiek*. Dit is nooit een statische checklist, maar moet worden begrepen als ‘ethiek als proces’: een voortdurende, iteratieve evaluatie waarbij niet alleen naar de code wordt gekeken, maar ook naar de organisatorische inbedding en de impact op de samenleving (Herzog & Blank, 2024; Bleher & Braun, 2023). Het gaat hierbij om het operationaliseren van waarden in de specifieke context van een algoritme. Waarbij men altijd waakzaam moet zijn voor gemakzuchtige ‘outsourcing’ van ethiek of 'ethics washing': het inzetten van ethiek om technologie slechts sneller geaccepteerd en uitgerold te krijgen. Onafhankelijke toetsing vervangt nooit de interne verantwoordelijkheid, maar ondersteunt en valideert deze. Het fungeert als een slot op de deur. Niet als een vervanging voor interne ethische verantwoordelijkheid of menselijk moreel besef binnen de eigen organisatie. (Herzog & Blank, 2024).

Context en Mensenrechten: Focus op live output

Een cruciaal element voor effectieve toetsing is de focus op *live output*, de daadwerkelijke beslissingen en interacties waarmee een burger wordt geconfronteerd. Een systeem kan in theorie accuraat zijn ('functional performance'), maar in de praktijk falen om effectief bij te dragen aan publieke doelen ('operational effectiveness') (Yeung & Li, 2025). Het is in de live output waar schendingen van mensenrechten zichtbaar worden en mensen geraakt worden. Instrumenten zoals het *Impact Assessment Mensenrechten en Algoritmes* (*IAMA*) van de Universiteit van Utrecht ‘dwingen’ tot een expliciete confrontatie tussen technologische mogelijkheden en fundamentele rechten (Gerards et al., 2026).

Conclusie

Onafhankelijke toetsing van publieke AI is geen luxe of bijzaak, maar een democratische noodzaak. Het vereist een verschuiving van abstracte principes naar concrete verantwoording over de live output in het publieke domein. Alleen door ethiek als een continu proces van onafhankelijke evaluatie te organiseren, kunnen we de kloof tussen theorie en praktijk dichten en waarborgen dat de digitale overheid de mensenrechten van haar burgers respecteert.

Literatuurlijst

Bengio, Y., Clare, S., Prunkl, C., Murray, M., Andriushchenko, M., Bucknall, B., ... & Mindermann, S. (2026). *International AI Safety Report 2026* (DSIT 2026/001). Department for Science, Innovation and Technology. https://internationalaisafetyreport.org

Bleher, H., & Braun, M. (2023). Reflections on putting AI ethics into practice: How three AI ethics approaches conceptualize theory and practice. *Science and Engineering Ethics*, *29*(3), 21. https://doi.org/10.1007/s11948-023-00443-3

Dignum, V. (2022). *Responsible artificial intelligence – from principles to practice*. Paper based on keynote at the Web Conference 2022. Umeå University.

Gerards, J., Muis, I., Straatman, J., Vankan, A., & Boiten, M. (2026). *IAMA Versie 2: Impact Assessment Mensenrechten en Algoritmes*. Universiteit Utrecht in opdracht van het Ministerie van Binnenlandse Zaken en Koninkrijksrelaties.

Herzog, C., & Blank, S. (2024). A systemic perspective on bridging the principles-to-practice gap in creating ethical artificial intelligence solutions – a critique of dominant narratives and proposal for a collaborative way forward. *Journal of Responsible Innovation*, *11*(1), 2431350. https://doi.org/10.1080/23299460.2024.2431350

OECD. (2024). *Governing with artificial intelligence: Are governments ready?* OECD Publishing.

Yeung, K., & Li, W. (2025). From ‘wild west’ to ‘responsible’ AI testing ‘in-the-wild’: Lessons from live facial recognition testing by law enforcement authorities in Europe. *Data & Policy*, *7*(e59). https://doi.org/10.1017/dap.2025.10019