Waarom deze checklist?

Een AI-agent in productie is geen demo. Hij leest documenten, mailt, boekt soms zelfs facturen. Eén foute instructie in een leveranciers-PDF kan een echte actie worden naar een echte klant. Deze checklist loopt zeven gebieden langs die wij afvinken voordat we een agent loslaten op productie-data.

Bedoeld voor agenten die mailboxen lezen, documenten van derden verwerken, externe websites ophalen, of acties initiëren met klantimpact. Niet voor zuivere FAQ-chatbots zonder schrijfrechten.

Vink eerlijk af. Wat openstaat is jouw eerste prioriteitenlijst.

1 Capability-mapping – wat mag deze agent eigenlijk?

Voor je iets dichttimmert, moet je weten wat er open staat.

- [ ] Lijst van alle tools en integraties die deze agent kan aanroepen (mail, ERP, CRM, scrapers, webhooks, file-storage).
- [ ] Per tool: lezen, schrijven, of beide?
- [ ] Per tool: scope (eigen klantcontext, eigen account, alle accounts)?
- [ ] Welke onomkeerbare acties zitten erbij? (Mail sturen, betaling initiëren, factuur boeken, bestelling plaatsen.)
- [ ] Wie heeft beslist dat de agent deze capabilities mag hebben? Datum, naam, beslissing genoteerd.

Resultaat: één pagina met de capability-matrix. Geen agent zonder deze pagina.

2 Scope-isolatie per stap

Eén grote agent met alle rechten is de standaard demo-opstelling – en de standaard manier om in de problemen te komen.

- [ ] Lees- en schrijf-stappen draaien als aparte rollen met aparte credentials.
- [ ] De stap die externe documenten leest heeft géén mail-, betaal- of mutatierechten.
- [ ] Tussen lezen en schrijven zit minimaal één gevalideerde overgang: een gestructureerd voorstel, geen vrije tekst die hergebruikt wordt als prompt.
- [ ] Per klant of tenant: aparte context. De agent ziet niet "alle klanten" tegelijk tenzij dat expliciet de bedoeling is.
- [ ] Credentials zijn niet hard-coded en niet gedeeld tussen rollen.

3 Allow-listing van tools per stap

Niet alleen welke tools de agent kan, maar welke per stap mogen.

- [ ] Per stap in de pipeline: expliciete lijst van toegestane tools.
- [ ] Geïnjecteerde tool-aanroepen worden door de runtime geweigerd, niet alleen door de modellogica.
- [ ] Log van geweigerde aanroepen wordt bijgehouden en steekproefsgewijs gereviewed.
- [ ] Geen wildcard tool-toegang in productie – dat is een demo-instelling.
- [ ] Tool-permissies zijn gekoppeld aan de pipeline-stap, niet aan de agent als geheel.

4 Output-validatie voor onomkeerbare acties

Het verschil tussen "de agent stelt voor om X te doen" en "de agent doet X".

- [ ] Onomkeerbare acties worden als gestructureerd voorstel geproduceerd, niet direct uitgevoerd.
- [ ] Validatieregels per actie-type: ontvanger op allow-list, bedrag binnen bandbreedte, taal-check op gevoelige inhoud.
- [ ] Hoog-risico-acties (bedrag boven drempel, nieuwe ontvanger, weekend of buiten kantooruren) escaleren naar mens-bevestiging.
- [ ] Rollback-pad beschreven: hoe draai je een foute actie terug, hoe snel, door wie?
- [ ] Audit-log per uitgevoerde actie, herleidbaar tot de stap en de input die hem opleverde.

5 Input-filters voor externe inhoud

Niet alle bronnen zijn even betrouwbaar — behandel ze ook zo.

- [ ] Bronnen geclassificeerd: vertrouwd (eigen documenten), semi-vertrouwd (vaste leveranciers), onvertrouwd (publieke web-scrapes, inkomende mail van onbekenden).
- [ ] OCR-tekst, metadata en verborgen tekstvelden (witte tekst, alt-attributen, document-comments) worden expliciet meegenomen in de scan.
- [ ] Bekende injection-patronen worden gedetecteerd en gemarkeerd ("negeer eerdere instructies", "system:"-prefixes, imperatieve commando's in onvertrouwde tekst).
- [ ] Bij injection-verdenking stopt de agent of escaleert, geen stille executie.
- [ ] Vertrouwen-niveau van de bron beïnvloedt welke tools daarna nog beschikbaar zijn.

6 Monitoring en escalatie

Een agent zonder logs is een agent die je niet kunt verdedigen.

- [ ] Elke tool-aanroep gelogd: tijdstip, input-hash, output, beslissing.
- [ ] Anomalie-detectie op tool-volgordes (een agent die normaal nooit mailt en plots wel).
- [ ] Hard limit op aantal acties per tijdsvenster — een gekaapte agent stopt automatisch.
- [ ] Escalatiepad bij verdacht gedrag: wie wordt gebeld, binnen welke termijn, met welke informatie?
- [ ] Dashboard of rapport waar wekelijks naar gekeken wordt — niet alleen "logs voor het geval dat".

7 Periodieke red-team check

Een keer per kwartaal: stop er bewust iets vervelends in.

- [ ] Test-document met verstopte instructie door de pipeline halen.
- [ ] Verifiëren dat de runtime de injection blokkeert en logt — niet alleen dat het model er niet in trapt.
- [ ] Nieuwe injection-technieken uit de OWASP LLM Top 10 en publieke incident-rapportages doornemen.
- [ ] Resultaten vastleggen, lessen verwerken in input-filters.
- [ ] Eigenaar van de red-team check benoemd. Geen anonieme verantwoordelijkheid.

Wat als je vastloopt?

Geen schande als deze checklist je laat zien dat er nu een paar zaken niet op orde zijn. Dat is de hele reden om hem af te vinken vóór een incident, niet erna.

Wij helpen MKB-bedrijven met dit type assessment op uurbasis. Vier tot acht uur, geen abonnement, geen verkoopgesprek. We kijken waar het wringt en geven je drie concrete prioriteiten. Daarna beslis je zelf wat je in huis doet en wat je uitbesteedt.

Contact

Spies Creations – Senior craft, AI-leverage voor MKB

Stephan Spies – 26 jaar IT-ondernemer

Website: spiescreations.nl

E-mail: contact@spiescreations.nl

Versie mei 2026. Laatste versie altijd via spiescreations.nl/downloads.