Ministerie van Onderwijs, Cultuur en Wetenschap

Van inzicht naar richting

Samengebrachte onderzoeksinzichten voor een gericht vervolgtraject.



Rijksprogramma
Duurzaam
Digitale
Informatiehuishouding

Versiebeheer	
30 april 2026	Versie 0.1, eerste opzet van het rapport.
26 mei 2026	Versie 0.2, verwerking feedback van projectteam.
22 juni 2026	Versie 0.3, verwerking feedback Nationaal Archief.
8 juli 2026	Versie 0.4, aangeboden aan het MT RDDI.
9 juli 2026	Versie 0.9, verwerking feedback MT RDDI.

Inhoud

Begrippenkader 4

1. Inleiding 5

1.1 Achtergrond 5
1.2 Doel van het rapport 5
1.3 Doelgroep 5
1.4 Leeswijzer 5

2. Onderzoeksaanpak 6

2.1 Doelstelling en centrale vraag 6
2.2 Deelvragen 6
2.3 Onderzoeksstrategie 6
2.4 Analysemethode 6
2.5 Afbakening en uitgangspunten 7

3. Rapporten 7

3.1 Machine Learning en automatische classificatie (Nationaal Archief, 2018) 7
3.2 Vooronderzoek e-mailrubricering (RDDI, 2025) 8
3.3 Automatische e-mailarchivering: doorontwikkeling van het Test Purposes Model (AZ, 2025) 8
3.4 Automatische e-mailclassificatie (RDDI, 2026) 8

4. Analyse 8

4.1 Dimensie: Techniek 9
4.2 Dimensie: Data 10
4.3 Dimensie: Rol van de mens 13
4.4 Dimensie: Juridisch 15
4.5 Dimensie: Organisatie en implementatie 17

5. Conclusies en aanbevelingen 20

5.1 Conclusie 20
5.2 Aanbevelingen 21

Bijlage 25

Begrippenkader

Begrip	Uitleg
Accuracy	Accuracy geeft het percentage correcte voorspellingen weer ten opzichte van het totaal aantal voorspellingen.
Artificial Intelligence (AI)	Artificial Intelligence is een verzamelnaam voor systemen of machines die taken kunnen uitvoeren waarvoor normaal gesproken menselijke intelligentie nodig is.
Betrouwbaarheidsdrempels	Betrouwbaarheidsdrempels (confidence thresholds) zijn vooraf gedefinieerde minimumscores, vaak variërend van 0 tot 1 of van 0% tot 100% die in machine learning worden gebruikt om te bepalen of de voorspelling van een model betrouwbaar genoeg is om op te reageren.
	Classificering is het labelen van e-mails op basis van de inhoud van een e-mail. De inhoud van de e-mail bepaalt welke label de e-mail krijgt.
Classificering	De confidence score is de waarschijnlijkheid die een model toekent aan een voorspelling en geeft aan hoe zeker het model is van een classificatie.
Confidence score	De drempelwaarde (threshold) is de grenswaarde die bepaalt vanaf welke confidence score een classificatie wordt geaccepteerd of automatisch wordt verwerkt.
Drempelwaarde	Dit zijn software-hulpmiddelen die de 'black box' van complexe AI doorbreken. Ze vertalen de interne logica van een AI-model naar een uitleg die voor mensen begrijpelijk is.
Explainability-methoden	Een werkwijze waarbij menselijke controle wordt ingezet bij twijfelgevallen om de kwaliteit en betrouwbaarheid van classificaties te waarborgen.
Human-in-the-loop	Machine learning is de techniek binnen AI die computers leert om zelfstandig patronen te ontdekken in data.
Machine Learning (ML)	Een PST-bestand (Personal Storage Table) is een Outlook-gegevensbestand (.pst) dat wordt gebruikt voor het opslaan van e-mails, agenda-items, contactpersonen en andere gegevens, vaak voor archivering of back-ups.
PST-bestand	Een PST-bestand (Personal Storage Table) is een Outlook-gegevensbestand (.pst) dat wordt gebruikt voor het opslaan van e-mails, agenda-items, contactpersonen en andere gegevens, vaak voor archivering of back-ups.

1. Inleiding

De afgelopen jaren zijn meerdere onderzoeken uitgevoerd naar automatische e-mailclassificatie en de inzet van AI binnen e-mailarchivering. Deze onderzoeken hebben waardevolle inzichten en rapporten opgeleverd, elk vanuit een eigen context, doelstelling en aanpak. Gezamenlijk bieden zij een rijk beeld van de mogelijkheden, beperkingen en aandachtspunten rondom dit vraagstuk.

1.1 Achtergrond

Tegelijkertijd maakt de variatie tussen onderzoeken het lastig om tot een eenduidig beeld te komen. De resultaten zijn verspreid over verschillende studies, waarbij accenten verschillen en conclusies niet altijd direct vergelijkbaar zijn. Hierdoor ontbreekt een integraal overzicht dat helpt om gefundeerde keuzes te maken voor de volgende fase.

1.2 Doel van het rapport

Om een goed vervolg te kunnen bepalen, is het van belang om de belangrijkste inzichten uit de bestaande onderzoeken te bundelen, te duiden en in samenhang te beschouwen. Door overeenkomsten en verschillen expliciet te maken, ontstaat beter zicht op wat als robuuste kennis kan worden beschouwd en waar nog onzekerheden of open vragen liggen.

Dit rapport voorziet in die behoefte door bestaande onderzoeksresultaten systematisch te analyseren en te synthetiseren. De opgedane inzichten worden vertaald naar een samenhangend perspectief en bieden daarmee een onderbouwde basis voor het bepalen van vervolgstappen. Daarbij ligt de nadruk op het identificeren van kansrijke richtingen voor verdere ontwikkeling en implementatie van AI-ondersteunde e-mailarchivering.

1.3 Doelgroep

Dit rapport richt zich op betrokkenen die verantwoordelijk zijn voor beleid, uitvoering en verdere ontwikkeling van AI-ondersteunde e-mailarchivering binnen de rijksoverheid. Het biedt hun een geconsolideerd beeld van de bestaande kennisbasis en ondersteunt bij het maken van onderbouwde keuzes voor een vervolgtraject.

1.4 Leeswijzer

Dit rapport is opgebouwd van context naar analyse en van analyse naar richting voor vervolg.

- Hoofdstuk 1 introduceert de aanleiding, doelgroep en doelstelling van het rapport.
- Hoofdstuk 2 beschrijft de onderzoeksaanpak, waaronder de centrale vraag, deelvragen, onderzoeksstrategie, analysemethode en afbakening.
- Hoofdstuk 3 geeft een beknopt overzicht van de vier onderzochte rapporten en plaatst deze in chronologische volgorde, zodat zichtbaar wordt hoe de inzichten zich in de tijd hebben ontwikkeld.
- Hoofdstuk 4 vormt de kern van het rapport. In dit hoofdstuk worden de onderzoeken geanalyseerd langs vijf dimensies: techniek, data, rol van de mens, juridisch en organisatie en implementatie. Per dimensie wordt ingegaan op de belangrijkste bevindingen, overeenkomsten en verschillen tussen de onderzoeken, onderliggende aannames, kennishiaten en kansrijke richtingen voor vervolg.
- Hoofdstuk 5 brengt de analyse samen in conclusies en aanbevelingen. Hier wordt beschreven wat de gezamenlijke inzichten betekenen voor een vervolgtraject en welke keuzes daarbij centraal staan.
- De bijlage bevat een beslismatrix waarin de vier onderzoeken langs de belangrijkste dimensies worden vergeleken.

2. Onderzoeksaanpak

In dit hoofdstuk wordt toegelicht hoe het onderzoek is opgezet en uitgevoerd. Achtereenvolgens wordt ingegaan op de doelstelling en centrale vraag, de deelvragen die richting geven aan de analyse, en de gehanteerde onderzoeksstrategie en analysemethode. Hiermee wordt inzicht gegeven in de wijze waarop de bestaande onderzoeken zijn geanalyseerd en hoe de conclusies in dit rapport tot stand zijn gekomen.

2.1 Doelstelling en centrale vraag

Het doel van dit onderzoek is om te komen tot een samenhangend en gedragen inzicht in de uitkomsten van bestaande onderzoeken naar AI-ondersteunde e-mailarchivering. Op basis hiervan wordt richting gegeven aan een mogelijk vervolgtraject. De centrale vraag die in dit onderzoek wordt beantwoord luidt:

Welke conclusies kunnen integraal worden getrokken uit bestaande onderzoeken naar AI-ondersteunde e-mailarchivering, en wat betekent dit voor een vervolgproject?

2.2 Deelvragen

Om de centrale vraag te beantwoorden, wordt ingezoomd op een aantal deelvragen. Allereerst wordt per onderzoek gekeken naar de belangrijkste bevindingen en de aannames die daaraan ten grondslag liggen. Vervolgens wordt geanalyseerd waar de verschillende onderzoeken elkaar versterken of juist tegenspreken. Daarnaast wordt expliciet gemaakt welke vragen nog onbeantwoord zijn gebleven en waar aanvullende kennis nodig is. Tot slot wordt verkend welke richtingen kansrijk zijn voor een vervolgtraject. De volgende deelvragen staan centraal:

- Wat zijn de belangrijkste bevindingen per onderzoek?
- Waar versterken de onderzoeken elkaar of spreken ze elkaar tegen?
- Welke aannames liggen ten grondslag aan de conclusies?
- Welke kennishiaten zijn er?
- Wat zijn kansrijke richtingen voor vervolg?

2.3 Onderzoeksstrategie

Het onderzoek heeft het karakter van een kwalitatieve synthese. De kern wordt gevormd door een systematische analyse van bestaande rapporten. Waar nodig wordt deze analyse aangevuld met duiding van betrokken experts. Deze aanpak sluit aan bij de aard van de opdracht: er wordt geen nieuw empirisch onderzoek uitgevoerd, maar bestaande kennis wordt samengebracht en geïnterpreteerd.

2.4 Analysemethode

Bij de analyse wordt gebruikgemaakt van principes uit thematische analyse en narratieve synthese. Dit betekent dat bevindingen niet alleen worden samengevat, maar ook worden geordend rond terugkerende thema's. Vervolgens worden deze thema's in samenhang geïnterpreteerd, waarbij aandacht is voor zowel overeenkomsten als verschillen tussen de onderzoeken.

2.5 Afbakening en uitgangspunten

In dit onderzoek wordt uitsluitend gebruikgemaakt van reeds uitgevoerde studies naar AI-ondersteunde e-mailarchivering binnen de rijksoverheid. De selectie van deze onderzoeken is bepalend voor de reikwijdte van de conclusies. Eventuele verschillen in context, doelstellingen en gehanteerde methoden tussen de oorspronkelijke studies worden meegenomen in de interpretatie van de resultaten.

Een bewuste afbakening binnen de selectie van studies is de keuze voor onderzoeken die direct toepasbaar zijn op de specifieke context van de Nederlandse overheid. De geselecteerde studies dienen hiermee te voldoen aan de randvoorwaarden die de Nederlandse wet- en regelgeving (in het bijzonder de Archiefwet en de AVG) stelt. Dit houdt onder andere in dat de onderzochte AI-toepassingen de authenticiteit, herleidbaarheid en transparantie van overheidsinformatie moeten borgen, evenals de bescherming van persoonsgegevens. Het correct classificeren van e-mails door AI mag er namelijk nooit toe leiden dat de context, bewijswaarde of reconstructie van bestuurlijke dossiers in het geding komt.

Deze focus op de Nederlandse context neemt niet weg dat onderzoeken van buitenlandse archiefinstellingen, onderwijsinstellingen of marktpartijen op het gebied van eDiscovery, e-mailarchivering en digitaal informatiebeheer eveneens waardevolle inzichten kunnen bevatten. Hoewel dergelijke studies buiten de reikwijdte dit onderzoek vallen, bieden zij een relevant referentiekader.

3. Rapporten

Dit hoofdstuk geeft een overzicht van de onderzoeken die als basis dienen voor de analyse in dit rapport. De rapporten worden in chronologische volgorde gepresenteerd, zodat inzicht ontstaat in de ontwikkeling van inzichten, aanpakken en toepassingen van AI-ondersteunde e-mailclassificatie binnen de rijksoverheid.

De geselecteerde rapporten laten een duidelijke ontwikkeling zien: van eerste experimenten en verkenningen naar meer toegepaste pilots en doorontwikkeling richting implementatie. Waar de eerdere onderzoeken zich met name richten op het verkennen van technische mogelijkheden en beperkingen, verschuift de focus in latere studies naar toepasbaarheid in de praktijk, inclusief juridische, organisatorische en beleidsmatige aspecten.

Dit hoofdstuk heeft als doel om de context en reikwijdte van de afzonderlijke onderzoeken te schetsen. Per rapport wordt kort ingegaan op de aanleiding, doelstelling en opzet. De inhoudelijke vergelijking en duiding van de bevindingen volgt in de daaropvolgende hoofdstukken.

3.1 Machine Learning en automatische classificatie (Nationaal Archief, 2018)

Focus van het rapport: Verkennen van de technische mogelijkheden en beperkingen van machine learning voor e-mailclassificatie (proof-of-concept).

Dit onderzoek betreft een experimentele verkenning naar de inzet van machine learning voor het classificeren van e-mails binnen de overheid. Het doel is om inzicht te krijgen in de vraag of zelflerende systemen in staat zijn om onderscheid te maken tussen functionele en niet-functionele e-mails, waarbij de laatste in het onderzoek wordt beschreven als niet-taak gerelateerd of ruis, en welke randvoorwaarden daarbij een rol spelen.

Het rapport beschrijft de ontwikkeling en toepassing van een prototype, waarbij medewerkers actief betrokken zijn bij het trainen van het model. Naast technische resultaten wordt ook aandacht besteed aan aspecten zoals vertrouwen, transparantie en de impact op werkprocessen. Hiermee vormt dit onderzoek een eerste stap in het verkennen van de mogelijkheden van AI binnen e-mailarchivering.

3.2 Vooronderzoek e-mailrubricering (RDDI, 2025)

Focus van het rapport: Aantonen van de potentie van automatische classificatie en de mate waarin deze de medewerker kan ondersteunen.

Dit vooronderzoek richt zich op de haalbaarheid van een technische oplossing die medewerkers ondersteunt bij het classificeren van e-mails in het kader van archivering. Daarbij staat de ontwikkeling van een ‘technisch assistent’ centraal die e-mails automatisch kan labelen of signaleren voor handmatige beoordeling.

Het onderzoek laat zien dat automatisering substantiële ondersteuning kan bieden, bijvoorbeeld door een groot deel van de privé-e-mails automatisch te filteren en de hoeveelheid te beoordelen e-mails te reduceren. Tegelijkertijd wordt benadrukt dat verdere ontwikkeling nodig is, met name op basis van realistische data en gebruikersinteractie.

3.3 Automatische e-mailarchivering: doorontwikkeling van het Test Purposes Model (AZ, 2025)

Focus van het rapport: Uitwerken van de functionele, technische en juridische eisen voor automatische e-mailarchivering (van concept naar randvoorwaarden voor implementatie).

Dit onderzoek richt zich op de doorontwikkeling van een Test Purposes Model (TPM) om inzicht te krijgen in de eisen en haalbaarheid van automatische e-mailarchivering binnen de rijksoverheid.

Het doel is niet alleen om de technische werking te toetsen, maar ook om de randvoorwaarden voor een werkbare en juridisch verantwoorde toepassing in kaart te brengen.

In het rapport wordt aandacht besteed aan classificatievraagstukken zoals bewaartermijnen, modelkeuzes en uitlegbaarheid, evenals aan de noodzaak van menselijke controle en feedback. Daarnaast worden belangrijke juridische en organisatorische aspecten benoemd, waaronder transparantie, reproduceerbaarheid en compliance met wet- en regelgeving.

3.4 Automatische e-mailclassificatie (RDDI, 2026)

Focus van het rapport: Integrale toetsing van automatische e-mailclassificatie in de praktijk (technisch, juridisch en organisatorisch).

Deze pilot onderzoekt de inzet van automatische e-mailclassificatie in een realistische praktijkcontext. Het doel is om te bepalen in hoeverre geautomatiseerde ondersteuning kan bijdragen aan het verbeteren van e-mailbeheer binnen de rijksoverheid, en onder welke voorwaarden dit verantwoord kan worden toegepast.

Het onderzoek combineert technische experimenten met juridische en organisatorische duiding. Daarbij wordt onder meer gekeken naar modelprestaties, datakwaliteit, classificatiecategorieën en de rol van menselijke controle. De resultaten laten zien dat automatische classificatie technisch haalbaar is en binnen de geldende kaders kan worden toegepast, mits zorgvuldig ingericht. Tegelijkertijd wordt duidelijk dat datakwaliteit, eenduidige definities en governance cruciale randvoorwaarden zijn voor succesvolle toepassing.

4. Analyse

Om de bevindingen uit de verschillende onderzoeken systematisch te kunnen vergelijken, is gebruikgemaakt van een analysematrix, zie Bijlage I. In deze matrix zijn de belangrijkste dimensies, waaronder technische haalbaarheid, datakwaliteit, rol van de mens, juridische randvoorwaarden en organisatorische implicaties, per rapport uitgewerkt.

De matrix maakt het mogelijk om overeenkomsten, verschillen en ontwikkelingen in de tijd inzichtelijk te maken. Op basis hiervan worden in dit hoofdstuk de belangrijkste patronen en inzichten beschreven. De analyse is thematisch opgebouwd langs de bovenstaande dimensies en gaat per dimensie in op de deelvragen reeds beschreven in 2.2.

4.1 Dimensie: Techniek

Binnen deze dimensie wordt gekeken naar de technische haalbaarheid van automatische e-mailclassificatie, de ontwikkeling van modelkeuzes en de voorwaarden waaronder techniek betrouwbaar kan worden toegepast.

4.1.1 Wat zijn de belangrijkste bevindingen per onderzoek?

De onderzochte rapporten laten een duidelijke ontwikkeling zien in de technische haalbaarheid van automatische e-mailclassificatie. In de vroege fase ligt de nadruk op experimentele verkenningen, waarbij wordt aangetoond dat machine learning in staat is om onderscheid te maken tussen typen e-mails (Nationaal Archief, 2018). In latere studies verschuift de focus naar praktische toepasbaarheid en worden modellen getest in meer realistische omgevingen (RDDI, 2025; AZ, 2025; RDDI, 2026).

De resultaten laten zien dat automatische classificatie technisch haalbaar is en in de praktijk toepasbaar kan zijn, met name wanneer deze wordt ingezet als ondersteuning van het classificatieproces (RDDI, 2025; RDDI, 2026). Tegelijkertijd wordt duidelijk dat prestaties variëren afhankelijk van de complexiteit van de classificatietaak. Eenvoudige onderscheidingen (zoals bewaren vs. vernietigen) blijken relatief goed automatiseerbaar, terwijl meer gedetailleerde classificaties aanzienlijk complexer zijn (RDDI, 2026).

Dit betekent dat technische haalbaarheid niet generiek is, maar taakafhankelijk: systemen werken betrouwbaar voor grofmazige beslissingen, maar verliezen nauwkeurigheid bij fijnmazige of contextgevoelige classificaties.

Daarnaast wordt zichtbaar dat verschillende modeltypen worden toegepast, variërend van basis machine learning-modellen tot meer geavanceerde technieken (AZ, 2025). Tegelijkertijd laten de onderzoeken zien dat deze verschillen in modelkeuze minder bepalend zijn voor de prestaties dan vaak wordt aangenomen, en dat andere factoren, zoals de aard van de taak en de context waarin deze wordt toegepast, een grotere rol spelen (RDDI, 2026).

De implicatie hiervan is dat de keuze voor een specifiek model niet doorslaggevend is. In de praktijk hangen de prestaties namelijk veel meer af van de data, definities en classificatiestructuur dan van de complexiteit van het algoritme.

4.1.2 Waar versterken de onderzoeken elkaar of spreken ze elkaar tegen?

De onderzoeken versterken elkaar in de conclusie dat automatische e-mailclassificatie technisch haalbaar is. Zowel in de experimentele fase als in latere praktijktoepassingen wordt aangetoond dat machine learning relevante patronen kan herkennen en classificaties kan ondersteunen (Nationaal Archief, 2018; RDDI, 2026).

Tegelijkertijd laten de onderzoeken een duidelijke ontwikkeling zien in de interpretatie van deze haalbaarheid. Waar vroege studies impliceren dat classificatie relatief eenvoudig is, tonen latere onderzoeken aan dat de techniek alleen robuust werkt onder specifieke voorwaarden (duidelijke categorieën, voldoende data, beperkte complexiteit) (RDDI, 2026).

Daarnaast verschillen de onderzoeken in hun focus en uitwerking. Het TPM-onderzoek richt zich expliciet op het vergelijken van verschillende modellen en het verkennen van technische keuzes (AZ, 2025), terwijl andere studies zich meer richten op toepasbaarheid in de praktijk (RDDI, 2025; RDDI, 2026).

Dit laat zien dat technische prestaties niet los kunnen worden gezien van toepassing: een model dat goed presteert in een experiment is niet automatisch geschikt voor implementatie. Hieruit ontstaat een spanningsveld tussen technische haalbaarheid en praktische toepasbaarheid: de techniek werkt in principe, maar niet in alle gevallen even goed, en vraagt om gerichte inzet afhankelijk van het type classificatieprobleem.

4.1.3 Welke aannames liggen ten grondslag aan de conclusies?

In de vroege onderzoeken wordt impliciet aangenomen dat e-mailclassificatie vergelijkbaar is met bestaande toepassingen zoals spamfilters (Nationaal Archief, 2018).

Deze aanname blijkt slechts gedeeltelijk geldig: e-mailclassificatie voor archivering is geen puur technisch patroonprobleem, maar een interpretatievraagstuk met context en juridische implicaties. Daarnaast wordt in meerdere onderzoeken verondersteld dat één modeltype breed toepasbaar is op verschillende classificatieproblemen. Het TPM-onderzoek laat echter zien dat verschillende modellen verschillende prestaties leveren afhankelijk van de toepassing, en dat modelkeuze niet los kan worden gezien van context en doel (AZ, 2025).

Ook ligt impliciet de aanname ten grondslag dat technische verbetering primair wordt bereikt door betere modellen. De onderzoeken laten zien dat deze aanname onvolledig is: verbeteringen in modelarchitectuur leveren minder op dan verbeteringen in data en definities.

4.1.4 Welke kennishiaten zijn er?

De analyse laat zien dat er nog beperkt inzicht is in de prestaties van modellen bij complexe en fijnmazige classificaties, met name in realistische werkomgevingen (RDDI, 2026). Cruciaal ontbreekt inzicht in foutverdeling: welke fouten worden gemaakt en wat de impact daarvan is op archivering.

Hoewel de technische haalbaarheid in algemene zin is aangetoond, is minder duidelijk onder welke voorwaarden en in welke contexten modellen betrouwbaar blijven functioneren.

Daarnaast bestaat er onzekerheid over de schaalbaarheid van technische oplossingen. Het TPM-onderzoek en de pilot laten zien dat implementatie mogelijk is, maar geven nog geen volledig beeld van prestaties op grotere schaal of bij langdurig gebruik (AZ, 2025; RDDI, 2026). Ook ontbreekt nog voldoende inzicht in de optimale combinatie van technieken, en in de vraag wanneer eenvoudige oplossingen volstaan en wanneer complexere modellen noodzakelijk zijn.

4.1.5 Wat zijn kansrijke richtingen voor vervolg?

Op basis van de analyse ontstaat een aantal duidelijke richtingen voor vervolg.

Ten eerste ligt het voor de hand om automatisering te richten op specifieke, goed afgebakende deelproblemen, in plaats van te streven naar volledige automatisering (RDDI, 2026). Dit sluit aan bij het inzicht dat niet alle classificaties even goed automatiseerbaar zijn.

Daarnaast blijkt dat een combinatie van technieken kansrijk is, waarbij eenvoudige rule-based oplossingen worden gecombineerd met machine learning, afhankelijk van de aard van het probleem (AZ, 2025; RDDI, 2026). Verder ligt er een duidelijke noodzaak voor iteratieve ontwikkeling, waarbij modellen stapsgewijs worden getest en verbeterd in praktijkcontexten, in plaats van uitsluitend in experimentele settings (RDDI, 2026).

Tot slot is het van belang om technische oplossingen sterker te koppelen aan de specifieke context waarin zij worden toegepast, zodat rekening wordt gehouden met variatie in e-mailgebruik, organisatie en classificatiebehoeften (RDDI, 2026). Dit betekent dat een generieke “alles-in-één” oplossing niet realistisch is; implementatie vraagt om gefaseerde en probleemgerichte inzet.

4.2 Dimensie: Data

Binnen deze dimensie staat de rol van data centraal, met aandacht voor datakwaliteit, representativiteit, labelconsistentie en de invloed daarvan op modelprestaties.

4.2.1 Wat zijn de belangrijkste bevindingen per onderzoek?

De vier onderzoeken laten een duidelijke ontwikkeling zien in hoe met data wordt omgegaan: van kleinschalige, gecontroleerde datasets naar grootschalige, realistische en complexe overheidsdata. Parallel daaraan verschuift het inzicht van “data als input” naar “data als bepalende factor voor prestaties”. Data zijn niet langer een randvoorwaarde, maar de primaire beperkende factor voor prestaties.

Het onderzoek van het Nationaal Archief (2018) werkt met een relatief kleine dataset. Van circa 3.500 handmatig gelabelde e-mails. De nadruk ligt op het aantonen van technische haalbaarheid, waarbij data vooral dienen als trainingsmateriaal. Tegelijkertijd wordt al zichtbaar dat het model afhankelijk is van menselijke correcties en iteratieve labeling, wat impliceert dat datakwaliteit en feedbackloops essentieel zijn voor verbetering.

Het RDDI-vooronderzoek (2025) maakt een fundamenteel andere keuze: vanwege privacybeperkingen wordt gewerkt met synthetische en synthetisch gelabelde data. De trainingsset bestaat uit e-mails uit de Enron-dataset die met behulp van GPT zijn gelabeld en vertaald, aangevuld met volledig synthetische e-mails voor ondervertegenwoordigde categorieën zoals partijpolitiek. De validatieset is handmatig samengesteld (137 e-mails). Deze aanpak maakt het mogelijk om snel te experimenteren en categorieën te modelleren die in bestaande datasets nauwelijks voorkomen. Tegelijkertijd wordt expliciet erkend dat deze data niet representatief zijn voor de praktijk en dat vervolgonderzoek met echte data noodzakelijk is.

Het AZ-onderzoek (2025) zet de stap naar praktijk. Er wordt gewerkt met e-mails uit 12 boxen van medewerkers, waarbij een subset (maximaal 500 per e-mailbox) handmatig is geannoteerd door archivarissen en privacy-experts. Dit onderzoek maakt zichtbaar dat werken met echte data aanzienlijk complexer is: e-mails zijn ambigu, categorieën overlappen en annotatoren verschillen in oordeel. De inter-annotator agreement van 69% onderstreept dat de data zelf geen eenduidige “ground truth” bevat.

De RDDI-pilot (2026) schaalt dit verder op en werkt met 13.709 handmatig geclassificeerde e-mails, uit mailboxen van bewindspersonen en bestuursraadleden. Deze dataset bevat vijf categorieën en laat duidelijk zien dat data de beperkende factor zijn voor modelprestaties. Grote categorieën (functioneel, niet-functioneel) presteren goed, terwijl kleine categorieën (privé, personeelsvertrouwelijk, partijpolitiek) structureel slechter scoren. Met slechts 37 voorbeelden voor partijpolitiek is betrouwbare classificatie niet mogelijk. Dit betekent dat gemiddelde prestaties (zoals accuracy) een vertekend beeld geven en risico’s in kleine categorieën maskeren. Daarnaast wordt zichtbaar dat circa 5% van de labels discussie oproept, wat wijst op inherente ambiguïteit in de data.

De rol van data verschuift van ondersteunend naar dominant. Waar het in 2018 vooral gaat om “voldoende data om te trainen”, laten de latere onderzoeken zien dat datakwaliteit, representativiteit, labelconsistentie en categorieverdeling de prestaties bepalen.

4.2.2 Waar versterken de onderzoeken elkaar of spreken ze elkaar tegen?

De onderzoeken versterken elkaar in de conclusie dat datakwaliteit essentieel is voor succesvolle toepassing van automatische e-mailclassificatie. In alle studies wordt zichtbaar dat modellen alleen goed presteren wanneer de data consistent, voldoende en representatief zijn.

Tegelijkertijd ontstaat er een duidelijke spanning tussen synthetische en echte data. Het RDDI-vooronderzoek (2025) behaalt relatief hoge prestaties op synthetische datasets, omdat deze intern consistent zijn en goed aansluiten op de gekozen definities. Het AZ-onderzoek (2025) en de RDDI-pilot (2026) laten zien dat prestaties op echte data lager en minder stabiel zijn, omdat echte e-mails ambigu zijn, context missen of meerdere interpretaties toelaten. Dit betekent dat resultaten op synthetische data niet zonder meer te vertalen zijn naar de praktijk.

Een tweede spanning betreft de verdeling van categorieën. In alle onderzoeken blijkt dat dominante categorieën beter worden geleerd dan kleine categorieën. In RDDI 2026 wordt dit expliciet zichtbaar: de grootste categorieën bepalen de overall accuracy, terwijl kleine categorieën achterblijven. Dit leidt tot een vertekend beeld wanneer alleen naar gemiddelde prestaties wordt gekeken.

Een derde spanning betreft labelconsistentie. Waar in RDDI (2025) labels grotendeels synthetisch en daarmee consistent zijn, laten AZ (2025) en RDDI (2026) zien dat menselijke annotatoren verschillen in interpretatie. Dit betekent dat de “juiste” classificatie niet altijd objectief vastligt, wat directe gevolgen heeft voor modeltraining en evaluatie.

De onderzoeken spreken elkaar niet tegen, maar maken samen zichtbaar dat prestaties sterk afhankelijk zijn van de aard van de data. Synthetische data maken experimenten mogelijk, maar echte data bepaalt de praktische haalbaarheid.

4.2.3 Welke aannames liggen ten grondslag aan de conclusies?

Een eerste aanname is dat data objectief en eenduidig gelabeld kan worden. De onderzoeken met echte data laten zien dat dit niet het geval is. Annotatoren verschillen in oordeel en categorieën overlappen inhoudelijk. De data bevatten dus geen absolute waarheid, maar interpretaties. In werkelijkheid is data interpretatief en afhankelijk van menselijke keuzes, wat directe impact heeft op modelprestaties.

Een tweede aanname is dat voldoende trainingsdata beschikbaar is voor alle categorieën. In de praktijk blijkt dat met name kleine of gevoelige categorieën ondervertegenwoordigd zijn. RDDI (2026) laat zien dat categorieën met weinig voorbeelden niet betrouwbaar te modelleren zijn, ongeacht de gekozen techniek.

Een derde aanname is dat synthetische data een goede proxy zijn voor echte data. RDDI (2025) laat zien dat synthetische data bruikbaar zijn voor exploratie, maar erkent dat validatie met echte data noodzakelijk is. De latere onderzoeken bevestigen dat synthetische datasets minder complex en minder ambigu zijn dan echte overheidsdata.

Een vierde aanname is dat meer data automatisch leidt tot betere prestaties. De onderzoeken laten zien dat niet alleen de hoeveelheid, maar vooral de kwaliteit en representativiteit van data bepalend zijn. Meer data van slechte kwaliteit of met inconsistente labels leidt niet automatisch tot betere modellen.

Een vijfde aanname is dat labels stabiel zijn over tijd. RDDI (2026) laat zien dat classificatierichtlijnen continu moeten worden aangepast en dat nieuwe e-mailtypen nieuwe interpretaties vereisen. De data zijn daarmee dynamisch, niet statisch.

4.2.4 Welke kennishiaten zijn er?

Het eerste kennishiaat betreft de kwaliteit en representativiteit van datasets. Er is nog geen dataset beschikbaar die representatief is voor de volle breedte van overheidsmailboxen, inclusief variatie tussen organisaties, functies en gebruikspatronen. Het tweede hiaat betreft labelconsistentie en definities. De onderzoeken tonen aan dat categorieën niet eenduidig zijn en dat annotatoren verschillen in interpretatie. Er ontbreekt een gestandaardiseerde en breed gedragen set definities die consistent kan worden toegepast.

Het derde hiaat betreft kleine en gevoelige categorieën. Voor categorieën zoals partijpolitiek en personeelsvertrouwelijk is onvoldoende data beschikbaar om betrouwbare modellen te trainen. Tegelijkertijd zijn dit vaak juist de categorieën met de grootste juridische en organisatorische impact. Het vierde hiaat betreft de relatie tussen datakwaliteit en modelprestaties. Hoewel duidelijk is dat data belangrijk zijn, is nog onvoldoende gekwantificeerd hoe verbeteringen in datakwaliteit zich vertalen naar prestatieverbetering. Het vijfde hiaat betreft data over tijd (drift). Er is nog geen inzicht in hoe veranderingen in e-mailgebruik, taalgebruik of organisatiecontext de prestaties van modellen beïnvloeden.

4.2.5 Wat zijn kansrijke richtingen voor vervolg?

De eerste en belangrijkste richting is investeren in datakwaliteit. Dit betekent: meer gelabelde data, betere representativiteit, expliciete kwaliteitscontroles en continue verbetering van datasets.

De tweede richting is standaardisatie van definities en labels. Door categorieën scherper te definiëren en richtlijnen te verduidelijken kan labelconsistentie worden verbeterd.

De derde richting is het verbeteren van annotatieprocessen. Dit omvat: training van annotatoren, gebruik van meerdere annotatoren en kwaliteitscontrole en consensusmechanismen. De vierde richting is een gerichte aanpak voor kleine categorieën.

Dit kan bijvoorbeeld door:

- Aanvullende dataverzameling
- Alternatieve methoden (bijv. regels of detectie op metadata)
- Prioritering van categorieën op basis van risico

De vijfde richting is het ontwikkelen van feedbackloops en continue dataverbetering. Data moet niet eenmalig worden verzameld, maar continu worden aangevuld en verbeterd op basis van gebruik.
De zesde richting is combinatie van data- en systeemontwerp. Data kan niet los worden gezien van de toepassing: keuzes in workflow, gebruikersinteractie en classificatiestructuur beïnvloeden direct de kwaliteit van de data.

De gezamenlijke conclusie uit de vier onderzoeken is dat data niet slechts een randvoorwaarde is, maar de bepalende factor voor het succes van automatische e-mailclassificatie. Niet modelkeuze, maar: datakwaliteit, labelconsistentie, categorieverdeling en representativiteit bepalen in de praktijk de prestaties. De kernverschuiving is daarmee: van “welk model werkt het best?” naar “welke data maakt betrouwbare classificatie mogelijk?”

4.3 Dimensie: Rol van de mens

Binnen deze dimensie wordt onderzocht hoe de rol van de mens verandert door automatische classificatie, en waarom menselijke beoordeling noodzakelijk blijft voor kwaliteit, interpretatie en controle.

4.3.1 Wat zijn de belangrijkste bevindingen per onderzoek?

De vier onderzoeken laten een duidelijke verschuiving zien in de rol van de mens: van actief classificeren naar ondersteunde beslisser binnen een hybride systeem, waarbij menselijke beoordeling essentieel blijft voor kwaliteit, interpretatie en legitimiteit.

In het Nationaal Archief (2018) staat de mens centraal in het leerproces van het model. De medewerker beoordeelt elke classificatie en corrigeert waar nodig. Deze correcties worden gebruikt om het model verder te trainen. De rol van de mens is hier dus primair: zonder menselijke feedback kan het systeem niet functioneren of verbeteren. De mens fungeert als bron van waarheid én als trainingsmechanisme.

In het RDDI-vooronderzoek (2025) verschuift de rol van de mens naar die van ondersteunde gebruiker. Het model wordt ingezet als assistent die suggesties doet, waarna de medewerker deze bevestigt of corrigeert. Belangrijk is dat het systeem expliciet rekening houdt met onzekerheid, bijvoorbeeld door een “Onzeker”-label toe te kennen bij lage confidence score. Daarmee ontstaat een eerste vorm van human-in-the-loop: de mens wordt gericht ingezet op de gevallen waar het model twijfelt.

Het AZ-onderzoek (2025) verdiept deze rol verder door expliciet aandacht te besteden aan onzekerheid, drempelwaarden en menselijke controle. Het systeem werkt met probabilistische uitkomsten, zodat twijfelgevallen kunnen worden doorgezet naar menselijke beoordeling. Tegelijkertijd wordt zichtbaar dat ook menselijke classificatie niet eenduidig is: archivarissen bereiken onderling slechts circa 69% overeenstemming. Dit betekent dat de mens niet alleen controlemechanisme is, maar ook zelf onderdeel van de onzekerheid in het systeem.

De RDDI-pilot (2026) laat zien hoe de rol van de mens door automatisering niet verdwijnt, maar verschuift van uitvoerder naar regisseur. De pilot maakt expliciet dat menselijke beoordeling essentieel blijft omdat classificatie niet volledig objectiveerbaar is, met name bij grensgevallen en gevoelige categorieën Tegelijkertijd wordt zichtbaar dat menselijke classificatie tijdsintensief is en dat schaalbaarheid een uitdaging vormt. Al met al laat de RDDI-pilot (2026) zien dat menselijke beoordeling essentieel blijft voor modeltraining, contextinterpretatie, het afhandelen van twijfelgevallen en het borgen van kwaliteit en legitimiteit.

4.3.2 Waar versterken de onderzoeken elkaar of spreken ze elkaar tegen?

De onderzoeken versterken elkaar in de conclusie dat de mens onmisbaar blijft in het classificatieproces. In alle studies is sprake van een vorm van human-in-the-loop, waarbij menselijke beoordeling nodig is voor training, validatie of controle.

Tegelijkertijd ontstaat er een duidelijke ontwikkeling in hoe de mens wordt ingezet:

- NA 2018: mens als primaire classifier en trainer
- RDDI 2025: mens als validator van modeloutput
- AZ 2025: mens als beslisser bij onzekerheid en risicovolle gevallen
- RDDI 2026: mens als onderdeel van een bredere socio-technische keten

De onderzoeken tonen dat automatisering menselijke arbeid niet vervangt, maar herverdeelt: minder bulkwerk, meer beoordeling van complexe gevallen. Er ontstaat ook een belangrijke spanning tussen menselijke betrouwbaarheid en modelprestaties. Waar modellen vaak worden beoordeeld op nauwkeurigheid, laten de onderzoeken zien dat menselijke classificatie zelf ook inconsistent is. In AZ 2025 wordt dit expliciet zichtbaar in de inter-annotator agreement. Dit betekent dat modellen niet worden vergeleken met een absolute waarheid, maar met menselijke interpretaties die variabel zijn.

Een tweede spanning betreft efficiëntie versus controle. Automatisering kan de werkdruk verlagen, maar vereist tegelijkertijd menselijke controle bij twijfelgevallen. Dit betekent dat de winst van automatisering afhankelijk is van hoe goed het systeem erin slaagt om de juiste gevallen aan de mens voor te leggen.

Een derde spanning betreft vertrouwen in het systeem. Hoewel niet altijd expliciet gemeten, komt in de onderzoeken naar voren dat gebruikers het systeem moeten begrijpen en vertrouwen om het effectief te gebruiken. Dit hangt samen met uitlegbaarheid en de manier waarop resultaten worden gepresenteerd.

De onderzoeken vullen elkaar aan in het inzicht dat de mens niet vervangen wordt, maar anders wordt ingezet. De kernvraag verschuift van “kan de mens worden vervangen?” naar “hoe kan de mens effectief samenwerken met het systeem?”

4.3.3 Welke aannames liggen ten grondslag aan de conclusies?

Een eerste aanname is dat de mens een betrouwbare en consistente beoordelaar is. De onderzoeken laten zien dat dit slechts gedeeltelijk waar is. Annotatoren verschillen in interpretatie, vooral bij grensgevallen. Dit betekent dat menselijke labels niet altijd een stabiele “ground truth” vormen. In werkelijkheid is menselijke beoordeling variabel, wat betekent dat modellen niet tegen een absolute waarheid worden getoetst.

Een tweede aanname is dat menselijke beoordeling schaalbaar is. In de praktijk blijkt dat handmatig classificeren tijdsintensief is en moeilijk op te schalen. De RDDI-pilot laat zien dat een groot deel van de inspanning zit in het labelen van data en het oplossen van interpretatieverschillen.

Een derde aanname is dat de mens optimaal gebruik maakt van modeloutput. In de praktijk bestaat het risico dat gebruikers modeluitkomsten klakkeloos overnemen (automation bias) of juist wantrouwen. Dit aspect wordt in de onderzoeken benoemd, maar nog niet empirisch onderzocht.

Een vierde aanname is dat menselijke controle voldoende is om fouten te voorkomen. De onderzoeken laten zien dat menselijke controle noodzakelijk is, maar niet perfect. Fouten blijven mogelijk, zeker bij complexe of ambigue gevallen.

Een vijfde aanname is dat de rol van de mens stabiel blijft. In werkelijkheid verandert deze rol door de introductie van AI-systemen: van uitvoerend naar controlerend, van individueel naar systemisch.

4.3.4 Welke kennishiaten zijn er?

Het eerste kennishiaat betreft de optimale taakverdeling tussen mens en systeem. Hoewel duidelijk is dat beide nodig zijn, is nog niet vastgesteld welke taken het best door mensen en welke door modellen kunnen worden uitgevoerd.

Het tweede hiaat betreft automation bias en gebruikersgedrag. Er is nog weinig empirisch inzicht in hoe gebruikers omgaan met modeloutput:

- Wanneer vertrouwen zij het systeem?
- Wanneer corrigeren zij het?
- Hoe beïnvloedt dit de kwaliteit van beslissingen?

Het derde hiaat betreft schaalbaarheid van menselijke inzet. Het labelen van data en beoordelen van twijfelgevallen vraagt veel capaciteit. Het is nog onduidelijk hoe dit efficiënt georganiseerd kan worden op grotere schaal.

Het vierde hiaat betreft kwaliteit van menselijke labels. Variatie tussen annotatoren maakt het lastig om consistente trainingsdata te creëren. Methoden om consensus te bereiken of verschillen te structureren zijn nog beperkt uitgewerkt.

Het vijfde hiaat betreft interactieontwerp. De manier waarop het systeem communiceert met de gebruiker (bijv. confidence scores, uitleg, labels) heeft direct invloed op de effectiviteit van menselijke besluitvorming, maar is nog onvoldoende onderzocht.

4.3.5 Wat zijn kansrijke richtingen voor vervolg?

De eerste richting is het ontwikkelen van een expliciete human-in-the-loop strategie, waarin bewust wordt bepaald welke beslissingen automatisch kunnen worden genomen, welke menselijke beoordeling vereisen en wanneer het systeem moet escaleren naar de gebruiker. Hiermee wordt de samenwerking tussen mens en systeem expliciet ontworpen in plaats van impliciet verondersteld.

De tweede richting is het verbeteren van annotatieprocessen. Dit vraagt om een meer systematische aanpak, waarin training van annotatoren, het gebruik van meerdere beoordelaars en het toepassen van consensusmechanismen worden gecombineerd met een continue aanscherping van definities. Op deze manier kan de kwaliteit en consistentie van trainingsdata structureel worden verhoogd.

De derde richting is het ontwerpen van effectieve gebruikersinterfaces. Daarbij gaat het niet alleen om functionaliteit, maar vooral om het ondersteunen van besluitvorming, bijvoorbeeld door het tonen van confidence scores, het geven van uitleg bij classificaties en het zichtbaar maken van onzekerheid. Dit stelt gebruikers in staat om beter geïnformeerde en bewustere keuzes te maken.

De vierde richting is het onderzoeken en mitigeren van automation bias. Dit vraagt om gerichte aandacht voor gebruikersgedrag, onder andere door experimenten uit te voeren, ontwerpkeuzes te maken die kritisch denken stimuleren en feedbackmechanismen in te bouwen die gebruikers actief betrekken bij de beoordeling van modeloutput.

De vijfde richting is het optimaliseren van taakverdeling en efficiëntie. In de praktijk betekent dit dat automatisering wordt ingezet voor bulkverwerking, terwijl de mens zich richt op uitzonderingen en risicovolle gevallen, waarbij prioritering plaatsvindt op basis van impact en onzekerheid.

De zesde richting is het versterken van multidisciplinaire samenwerking. De pilot laat zien dat classificatievraagstukken niet uitsluitend technisch zijn, maar ook juridische, organisatorische en inhoudelijke dimensies hebben. Dit vraagt om een inrichting waarin deze perspectieven structureel samenkomen.

De gezamenlijke conclusie is dat de mens geen randvoorwaarde is, maar een integraal onderdeel van het systeem. Automatische e-mailclassificatie moet worden begrepen als een socio-technisch systeem waarin modellen ondersteunen, mensen interpreteren en controleren, en de interactie tussen beide bepaalt hoe effectief het systeem functioneert. De kernverschuiving is daarmee: van “mens versus machine” naar “mens en machine als samenwerkend systeem”.

4.4 Dimensie: Juridisch

Binnen deze dimensie wordt gekeken naar de juridische randvoorwaarden voor automatische e-mailclassificatie, waaronder privacy, transparantie, uitlegbaarheid, proportionaliteit en verantwoordelijkheid.

4.4.1 Wat zijn de belangrijkste bevindingen per onderzoek?

De vier onderzoeken laten een duidelijke ontwikkeling zien in de juridische duiding van automatische e-mailclassificatie: van impliciete randvoorwaarde naar expliciet integraal onderdeel van de beoordeling. Waar de vroege studies zich primair richten op technische haalbaarheid, verschuift de focus in latere onderzoeken naar de vraag onder welke juridische voorwaarden toepassing mogelijk is.

In het Nationaal Archief (2018) speelt het juridische perspectief een beperkte rol. Privacy en informatiebeveiliging worden benoemd als aandachtspunten, maar niet systematisch uitgewerkt. De nadruk ligt op technische verkenning, waarbij impliciet wordt aangenomen dat juridische vraagstukken oplosbaar zijn binnen bestaande kaders.

Het RDDI-vooronderzoek (2025) maakt een expliciete keuze om niet met echte overheidsdata te werken vanwege privacyrisico's. Dit is een belangrijke juridische ingreep: door te werken met synthetische data wordt voorkomen dat persoonsgegevens worden verwerkt in een experimentele setting. Tegelijkertijd betekent deze keuze dat de juridische haalbaarheid in de praktijk nog niet wordt aangetoond.

Het AZ-onderzoek (2025) brengt de juridische dimensie nadrukkelijk naar voren door automatische e-mailclassificatie niet alleen technisch maar ook juridisch te beoordelen. Hierbij wordt specifiek gekeken naar de AVG, de uitlegbaarheid van beslissingen, proportionaliteit en subsidiariteit, en de risico's van foutieve classificatie. Een belangrijk inzicht is dat probabilistische modellen noodzakelijk zijn om juridische controle mogelijk te maken: door te werken met zekerheidspercentages kunnen twijfelgevallen worden geïdentificeerd en aan menselijke beoordeling worden voorgelegd.

De RDDI-pilot (2026) integreert de juridische dimensie volledig in de analyse. De centrale vraag is niet alleen of automatische classificatie technisch mogelijk is, maar ook of deze juridisch toelaatbaar is. De pilot laat zien dat automatische classificatie binnen de kaders van de Archiefwet, Woo en AVG mogelijk is, mits zorgvuldig ingericht. Daarbij worden randvoorwaarden benoemd zoals: transparantie van het systeem, menselijke controle, expliciete governance, zorgvuldige omgang met persoonsgegevens.

De pilot concludeert dat juridische randvoorwaarden geen blokkade vormen, maar wel bepalend zijn voor de inrichting van het systeem.

Hiermee verschuift de juridische dimensie van randvoorwaarde naar ontwerpprincipe. De vraag is niet meer of het juridisch kan, maar hoe het juridisch verantwoord moet worden ingericht.

4.4.2 Waar versterken de onderzoeken elkaar of spreken ze elkaar tegen?

De onderzoeken versterken elkaar in de conclusie dat automatische e-mailclassificatie juridisch mogelijk is, mits aan bepaalde voorwaarden wordt voldaan. Er is geen studie die concludeert dat de techniek principieel onverenigbaar is met wet- en regelgeving. Tegelijkertijd laten de onderzoeken een duidelijke ontwikkeling zien in de mate van juridische explicitering:

- NA 2018: impliciet
- RDDI 2025: vermijdend (synthetische data)
- AZ 2025: expliciet juridisch kader
- RDDI 2026: geïntegreerde juridische beoordeling

Een belangrijke aanvulling tussen de onderzoeken betreft de rol van transparantie en uitlegbaarheid. AZ 2025 benadrukt dat systemen moeten kunnen uitleggen hoe een classificatie tot stand komt, met name vanwege de juridische consequenties van beslissingen. RDDI 2026 bevestigt dit en koppelt het aan governance en verantwoording.

Er ontstaat ook een spanning tussen automatisering en juridische controle. Enerzijds maakt automatisering het mogelijk om grote hoeveelheden e-mails efficiënt te verwerken. Anderzijds vereist de wetgeving dat beslissingen controleerbaar, uitlegbaar en proportioneel zijn. Dit leidt tot de noodzaak van menselijke controle bij twijfelgevallen.

Een tweede spanning betreft privacy versus datagebruik. RDDI 2025 vermijdt privacy risico's door synthetische data te gebruiken, maar dit beperkt de realiteitswaarde van de resultaten. AZ 2025 en RDDI 2026 werken met echte data en laten zien dat verwerking van persoonsgegevens mogelijk is, mits beveiliging, governance en rechtmatigheid goed zijn ingericht.

De onderzoeken vullen elkaar aan door te laten zien dat juridische haalbaarheid niet vanzelfsprekend is, maar afhankelijk van ontwerpkeuzes. Automatisering en juridische borging moeten in balans worden gebracht.

4.4.3 Welke aannames liggen ten grondslag aan de conclusies?

Een eerste aanname is dat juridische kaders een belemmering vormen voor automatisering. De latere onderzoeken laten zien dat dit niet per se het geval is. Juridische kaders zijn vooral richtinggevend: zij bepalen hoe een systeem moet worden ingericht, maar sluiten automatisering niet uit.

Een tweede aanname is dat transparantie en uitlegbaarheid vanzelf volgen uit technische implementatie. In de praktijk blijkt dat dit expliciet moet worden ontworpen. Complexe modellen zoals BERT zijn minder uitlegbaar, wat juridische risico's kan vergroten.

Een derde aanname is dat automatische classificatie juridisch gelijkwaardig is aan menselijke classificatie. De onderzoeken laten zien dat dit niet het geval is. Automatische beslissingen vereisen aanvullende waarborgen, zoals:

- Controlemechanismen
- Logging en auditability
- Expliciete drempelwaarden

Een vierde aanname is dat privacyvraagstukken oplosbaar zijn door data te anonimiseren of te vermijden. RDDI (2025) laat zien dat dit mogelijk is in een experimentele fase, maar dat toepassing in de praktijk juist vraagt om verwerking van echte data binnen juridische kaders.

Een vijfde aanname is dat menselijke controle voldoende is om juridische risico's te mitigeren. Hoewel human-in-the-loop essentieel is, blijft de vraag hoe deze controle effectief en schaalbaar kan worden ingericht.

4.4.4 Welke kennishiaten zijn er?

Het eerste kennishiaat betreft de praktische implementatie van uitlegbaarheid. Hoewel het belang hiervan wordt erkend, is nog onvoldoende duidelijk hoe complexe modellen juridisch uitlegbaar kunnen worden gemaakt in een productieomgeving.

Het tweede hiaat betreft de risicomodellering van fouten. Niet alle fouten zijn juridisch gelijk: het onterecht vernietigen van archiefwaardige e-mails is onomkeerbaar, terwijl het onterecht bewaren van privégegevens privacy risico's oplevert. Dit betekent dat generieke prestatiemaatstaven zoals accuracy onvoldoende zijn om de juridische impact te beoordelen. Er ontbreekt een systematische koppeling tussen fouttypen en juridische consequenties, waardoor onduidelijk blijft welke foutmarges acceptabel zijn en onder welke voorwaarden automatische classificatie verantwoord kan worden ingezet.

Het derde hiaat betreft governance en verantwoordelijkheid. Wie is verantwoordelijk voor een foutieve classificatie? De ontwikkelaar van het model? De organisatie? of de gebruiker? Deze vraag is nog onvoldoende uitgewerkt.

Het vierde hiaat betreft toepassing op schaal. De onderzoeken laten zien dat toepassing mogelijk is in pilots, maar er is nog weinig inzicht in juridische implicaties bij grootschalige, rijksbrede implementatie.

Het vijfde hiaat betreft interactie tussen wetgeving en techniek over tijd. Wetgeving (zoals AI Act) ontwikkelt zich, en het is nog onduidelijk hoe toekomstige regelgeving de toepassing van dergelijke systemen zal beïnvloeden.

4.4.5 Wat zijn kansrijke richtingen voor vervolg?

De eerste richting is het toepassen van compliance-by-design. Juridische randvoorwaarden moeten vanaf het begin worden meegenomen in het ontwerp van het systeem, niet achteraf. De tweede richting is het versterken van transparantie en uitlegbaarheid. Dit kan door:

- Gebruik van interpreteerbare modellen waar mogelijk
- Toevoegen van explainability-methoden (die bijvoorbeeld de labelkeuze van het model toe kan lichten)
- Inzicht geven in beslisregels en confidence scores

De derde richting is het ontwikkelen van duidelijke governance-structuren. Dit omvat: rolverdeling en verantwoordelijkheden, toezicht en auditing en documentatie van beslissingen.

De vierde richting is het implementeren van risicogestuurde classificatie. Hierbij worden drempelwaarden afgestemd op juridische risico's en gevoelige categorieën extra gecontroleerd.

De vijfde richting is het versterken van privacy- en beveiligingsmaatregelen. Dit vraagt om veilige dataomgevingen, DPIA's en expliciete afspraken over datagebruik.

De zesde richting is het integreren van menselijke controle als juridisch mechanisme. Human-in-the-loop moet niet alleen technisch, maar ook juridisch worden gepositioneerd als waarborg.

De gezamenlijke conclusie is dat automatische e-mailclassificatie juridisch mogelijk is, maar alleen binnen een zorgvuldig ontworpen kader. De kernverschuiving is: van "mag het?" naar "hoe moet het ingericht worden om te mogen?" De juridische dimensie fungeert daarmee niet als beperking, maar als ontwerpprincipe, kwaliteitskader en legitimatie voor toepassing.

4.5 Dimensie: Organisatie en implementatie

Binnen deze dimensie staat centraal hoe automatische e-mailclassificatie kan worden ingebed in werkprocessen, governance, beheer en organisatorische besluitvorming.

4.5.1 Wat zijn de belangrijkste bevindingen per onderzoek?

De vier onderzoeken laten een duidelijke ontwikkeling zien van technische experimenten naar de vraag hoe automatische e-mailclassificatie organisatorisch kan worden ingebed. Waar de vroege studies zich richten op wat technisch mogelijk is, verschuift de focus in latere onderzoeken naar de praktische inrichting van processen, systemen en governance.

In het Nationaal Archief (2018) wordt al impliciet zichtbaar dat implementatie meer vraagt dan alleen een werkend model. Het rapport benadrukt dat verschillende componenten model, infrastructuur, beveiliging en werkprocessen op elkaar moeten aansluiten. De oplossing wordt niet gepresenteerd als een losstaand algoritme, maar als onderdeel van een bredere systeemarchitectuur waarin de medewerker actief betrokken blijft.

Het RDDI-vooronderzoek (2025) verkent voor het eerst expliciet de rol van een technisch assistent in de werkpraktijk. De focus ligt op ondersteuning van medewerkers bij het classificeren van e-mails, met als doel het verminderen van werkdruk en het verbeteren van efficiëntie. Het rapport schetst een scenario waarin het model geïntegreerd wordt in de werkomgeving (bijvoorbeeld e-mailsoftware), en waarin medewerkers alleen nog de door het systeem geselecteerde of onzekere gevallen hoeven te beoordelen. Tegelijkertijd blijft de implementatie in deze fase conceptueel: er wordt nog niet getest in een echte werkomgeving.

Het AZ-onderzoek (2025) brengt implementatievraagstukken expliciet in beeld. Het Test Purposes Model wordt niet alleen technisch getest, maar ook geplaatst in de context van archiveringsprocessen. Het rapport beschrijft een mogelijke systeemarchitectuur waarin:

- Classificatie plaatsvindt in meerdere stappen (bijv. eerst bewaren/vernietigen, daarna verfijning)
- Onzekerheden worden doorgestuurd naar menselijke beoordeling
- Feedback wordt gebruikt voor modelverbetering
- Monitoring en governance onderdeel zijn van het systeem

Daarmee wordt duidelijk dat implementatie niet alleen een technische configuratie is, maar het resultaat van expliciete ontwerpkeuzes rondom processen, drempelwaarden en verantwoordelijkheidsverdeling.

De RDDI-pilot (2026) brengt deze inzichten samen in een praktijkcontext. Hier wordt duidelijk dat implementatie de meest complexe stap is. De pilot laat zien dat:

- Data-inwinning (bijv. uit PST-bestanden) technisch en organisatorisch complex is
- Datasets opgeschoond moeten worden (corruptie, duplicaten, niet-e-mailitems)
- Classificatierichtlijnen moeten worden vastgesteld en continu bijgesteld
- Samenwerking nodig is tussen verschillende rollen (data scientists, archivarissen, beleidsmakers)

Daarnaast wordt zichtbaar dat implementatie niet alleen een technische integratie is, maar een veranderproces. De manier waarop medewerkers werken, beslissingen nemen en omgaan met informatie verandert door de introductie van het systeem.

Implementatie verschuift van een technische stap naar een organisatorische transformatie. Het succes van automatische classificatie wordt niet bepaald door het model alleen, maar door de manier waarop het wordt ingebed in processen, systemen en rollen.

4.5.2 Waar versterken de onderzoeken elkaar of spreken ze elkaar tegen?

De onderzoeken versterken elkaar in de conclusie dat implementatie niet triviaal is en dat technische haalbaarheid niet automatisch leidt tot succesvolle toepassing. Automatische classificatie heeft alleen waarde wanneer deze wordt geïntegreerd in bestaande werkprocessen. Tegelijkertijd laten latere studies zien dat deze integratie ingrijpender is dan aanvankelijk werd verondersteld. Een duidelijke ontwikkeling is zichtbaar in de mate van concretisering:

- NA 2018: implementatie impliciet en technisch
- RDDI 2025: implementatie als concept (technisch assistent)
- AZ 2025: implementatie als systeemarchitectuur en procesontwerp
- RDDI 2026: implementatie als praktijkervaring met concrete knelpunten

Dit laat zien dat implementatie verschuift van een technisch vraagstuk naar een organisatievraagstuk.

Er ontstaat een spanning tussen efficiëntie en complexiteit. Enerzijds kan automatisering werkdruk verminderen, bijvoorbeeld door filtering van niet-relevante e-mails. Anderzijds introduceert het systeem nieuwe complexiteit, zoals modelbeheer, inrichting van drempelwaarden, monitoring en afstemming tussen rollen. Hierdoor is efficiëntie winst niet vanzelfsprekend, maar afhankelijk van de inrichting van het systeem.

Een tweede spanning betreft standaardisatie versus contextafhankelijkheid. Implementatie vraagt om gestandaardiseerde processen en definities, terwijl classificatie in de praktijk sterk afhankelijk is van context en gebruikssituatie. Dit beperkt de mogelijkheid tot een generieke oplossing. Een derde spanning betreft centralisatie versus decentrale toepassing. Hoewel de techniek generiek lijkt, verschillen mailboxen, werkprocessen en classificatiebehoeften per organisatie. Dit betekent dat implementatie altijd maatwerk vereist.

Implementatie is daarmee geen lineair proces van “modelbouwen” naar “uitrollen”, maar een iteratief proces waarin techniek, organisatie en context voortdurend op elkaar moeten worden afgestemd. De implicatie is dat succesvolle toepassing niet alleen vraagt om een goed model, maar om continue organisatorische inbedding en bijsturing.

4.5.3 Welke aannames liggen ten grondslag aan de conclusies?

Een eerste aanname is dat technische oplossingen eenvoudig in bestaande processen kunnen worden geïntegreerd. De onderzoeken laten zien dat dit niet het geval is. Implementatie vereist aanpassing van werkprocessen, rollen en verantwoordelijkheden.

Een tweede aanname betreft de veronderstelling dat automatisering direct leidt tot efficiëntiewinst. In de praktijk blijkt deze winst echter afhankelijk van meerdere factoren, zoals de kwaliteit van de output, de mate waarin menselijke controle noodzakelijk blijft en de wijze waarop processen zijn ingericht en ingebed binnen de organisatie. Daarnaast kan de complexiteit van het systeem er in sommige gevallen toe leiden dat een deel van de beoogde efficiëntiewinst verloren gaat.

Een derde aanname is dat één generieke oplossing toepasbaar is op alle organisaties. De onderzoeken laten zien dat implementatie sterk afhankelijk is van context, zoals: type organisatie, e-mailgebruik en archiveringspraktijken.

Een vierde aanname is dat governance vanzelf volgt uit implementatie. In de praktijk blijkt dat governance expliciet moet worden ingericht, inclusief: verantwoordelijkheden, controlemechanismen en besluitvorming.

Een vijfde aanname is dat gebruikers zich automatisch aanpassen aan het systeem. Implementatie blijkt echter een veranderproces te zijn dat vraagt om: training, begeleiding en acceptatie.

4.5.4 Welke kennishiaten zijn er?

Het eerste kennishiaat betreft de inrichting van werkprocessen. Er is nog geen uitgewerkt model voor hoe automatische classificatie optimaal wordt geïntegreerd in dagelijkse werkzaamheden. Het tweede hiaat betreft governance en verantwoordelijkheden. Hoewel het belang wordt erkend, is nog niet uitgewerkt:

- Wie eigenaar is van het systeem
- Wie verantwoordelijk is voor fouten
- Hoe toezicht wordt ingericht

Het derde hiaat betreft opschaling en beheer. De onderzoeken zijn uitgevoerd in beperkte contexten. Er is nog weinig inzicht in beheer van modellen op grote schaal, onderhoud van datasets en monitoring over langere tijd.

Het vierde hiaat betreft veranderimpact op organisaties. De introductie van AI-systemen verandert rollen en verantwoordelijkheden, maar deze impact is nog niet systematisch onderzocht.

Het vijfde hiaat betreft integratie met bestaande systemen. Technische integratie (bijv. met e-mailclients of archiefsystemen) wordt benoemd, maar nog niet volledig uitgewerkt.

4.5.5 Wat zijn kansrijke richtingen voor vervolg?

De eerste richting is het ontwikkelen van een duidelijk implementatiemodel, waarin procesinrichting, rolverdeling en interactie tussen mens en systeem expliciet worden vastgelegd. Zonder deze explicitering blijft onduidelijk hoe het systeem in de praktijk moet functioneren.

De tweede richting is het inrichten van governance-structuren, inclusief duidelijke verantwoordelijkheden, toezicht, auditing en besluitvorming over classificatie en drempelwaarden. Dit is noodzakelijk om controle en verantwoording te borgen in een geautomatiseerd proces.

De derde richting is gefaseerde implementatie: starten met beperkte toepassingen (zoals filtering) en vervolgens uitbreiden naar complexere taken. Dit verkleint risico's en maakt het mogelijk om stapsgewijs te leren en bij te sturen.

De vierde richting is het versterken van monitoring en beheer, met continue evaluatie van prestaties, detectie van afwijkingen (drift) en bijsturing van modellen en processen. Hiermee wordt voorkomen dat prestaties in de tijd ongemerkt verslechteren.

De vijfde richting is investeren in organisatieontwikkeling, waaronder training van medewerkers, begeleiding bij verandering en het stimuleren van acceptatie. Zonder deze investering blijft het systeem onderbenut of verkeerd toegepast.

De zesde richting is het ontwikkelen van flexibele en contextgevoelige oplossingen. In plaats van één standaardoplossing is maatwerk per organisatie nodig. Dit volgt uit de variatie in werkprocessen, data en classificatiebehoeften.

De gezamenlijke conclusie is dat implementatie de meest complexe en bepalende stap is in de toepassing van automatische e-mailclassificatie. De kernverschuiving is: van "kan het systeem werken?" naar "kan het systeem landen in de organisatie?". Succes wordt daarmee primair bepaald door organisatorische inbedding, niet door technische prestaties alleen.

5. Conclusies en aanbevelingen

5.1 Conclusie

De vier onderzochte rapporten laten gezamenlijk een duidelijke ontwikkeling zien: van experimentele verkenningen naar praktijkgerichte toepassingen van automatische e-mailclassificatie. Op basis hiervan kunnen enkele robuuste conclusies worden getrokken over de haalbaarheid en inzet van AI-ondersteunde e-mailarchivering.

Allereerst blijkt automatische e-mailclassificatie technisch haalbaar en in staat om waarde toe te voegen binnen de praktijk. Met name bij grofmazige classificaties en ter ondersteuning van medewerkers levert automatisering efficiëntievoordelen op. Tegelijkertijd is deze haalbaarheid niet generiek toepasbaar. De prestaties van modellen zijn sterk afhankelijk van factoren zoals de complexiteit van de classificatietaak, de kwaliteit van de data en de context waarin het systeem wordt ingezet. Vooral complexe, contextafhankelijke en kleinere categorieën blijven lastig te classificeren, waardoor volledige automatisering niet realistisch lijkt.

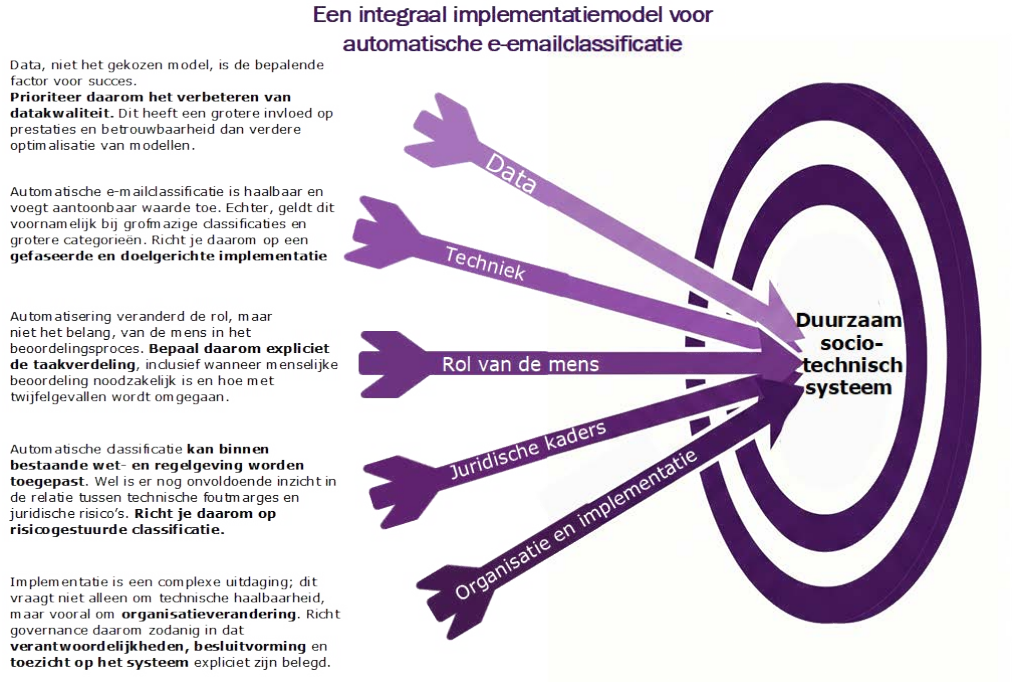
Daarnaast komt uit alle onderzoeken naar voren dat data de bepalende succesfactor vormt. Niet zozeer de keuze voor een specifiek model, maar de kwaliteit, representativiteit en consistentie van data blijken doorslaggevend voor de prestaties van het systeem. Hoewel synthetische data bruikbaar is voor verkennende doeleinden, biedt alleen realistische data inzicht in de daadwerkelijke toepasbaarheid. Juist deze data blijken in de praktijk echter vaak ambigu en inconsistent, waardoor een eenduidige “ground truth” ontbreekt.

Verder blijkt dat de rol van de mens essentieel blijft binnen het classificatieproces. Automatisering vervangt menselijke beoordeling niet, maar verandert de aard ervan. De rol verschuift van uitvoerende classificatie naar het beoordelen van complexe of risicovolle gevallen en het monitoren van de kwaliteit van het systeem. Effectieve toepassing vraagt daarom om een goed ingerichte samenwerking tussen mens en technologie.

Ook de juridische dimensie blijkt vooral richtinggevend in plaats van beperkend. Automatische classificatie kan binnen bestaande wet- en regelgeving worden toegepast, mits transparantie, controleerbaarheid en proportionaliteit expliciet worden geborgd. Tegelijkertijd ontbreekt nog voldoende inzicht in de relatie tussen technische foutmarges en juridische risico’s, waardoor het lastig blijft om acceptabele foutpercentages vast te stellen.

Tot slot laten de rapporten zien dat implementatie de meest complexe factor vormt. Technische haalbaarheid leidt niet automatisch tot succesvolle toepassing. Implementatie vraagt om organisatorische inbedding, aangepaste processen, duidelijke governance en acceptatie door gebruikers. Daarmee is automatische e-mailclassificatie niet uitsluitend een technische ontwikkeling, maar vooral een organisatieverandering.

Over alle dimensies heen ontstaat één samenhangend beeld: automatische e-mailclassificatie is geen op zichzelf staande technologie, maar een socio-technisch systeem waarin techniek, data, mens, juridische kaders en organisatie onlosmakelijk met elkaar verbonden zijn. De centrale vraag verschuift daarmee van “kan het systeem werken?” naar “onder welke voorwaarden kan het systeem verantwoord en effectief worden toegepast?”.



Figuur 1: Integraal implementatiemodel voor automatische e-mailclassificatie

5.2 Aanbevelingen

Op basis van de analyse volgen zes samenhangende aanbevelingen voor een vervolgtraject.

1. Gefaseerde en doelgerichte implementatie
Richt de implementatie gefaseerd in, waarbij gestart wordt met afgebakende toepassingen en stapsgewijs wordt uitgebreid naar complexere taken. Dit maakt het mogelijk om risico's beheersbaar te houden en keuzes gaandeweg te onderbouwen.

2. Focus op datakwaliteit
Geef prioriteit aan het verbeteren van datakwaliteit, waaronder representativiteit, labelconsistentie en heldere categoriedefinities. Verbeteringen op dit vlak hebben een grotere invloed op prestaties en betrouwbaarheid dan verdere optimalisatie van modellen.

3. Taakverdeling mens–systeem
Bepaal expliciet de taakverdeling tussen systeem en mens, inclusief wanneer menselijke beoordeling noodzakelijk is en hoe met twijfelgevallen wordt omgegaan. Dit zorgt voor een evenwichtige inzet van automatisering en voorkomt zowel overbelasting als onderbenutting van capaciteit.

4. Governance en verantwoordelijkheid
Richt governance zodanig in dat verantwoordelijkheden, besluitvorming en toezicht op het systeem expliciet zijn belegd. Hiermee wordt geborgd dat controle, uitlegbaarheid en verantwoording structureel zijn ingericht.

5. Risico gestuurde classificatie
Differentieer de inzet van automatisering op basis van risico per categorie en fouttype. Hierdoor kunnen strengere waarborgen worden toegepast waar de impact groot is, en kan efficiënter worden geautomatiseerd waar risico's beperkt zijn.

6. Integraal implementatiemodel
Ontwikkel een rijksbreed implementatiemodel waarin techniek, data, mens, juridische eisen en organisatie vanaf het begin samenhangend worden ontworpen. Dit voorkomt fragmentatie en vergroot de kans dat het systeem in de praktijk goed functioneert.

Het model moet bestaan uit een centrale basis voor alle ministeries, met gezamenlijke uitgangspunten, standaarden, juridische randvoorwaarden, governance en minimale implementatie-eisen. Tegelijk moet er ruimte zijn voor ministerie-specifieke toepassing, omdat ministeries verschillen in processen, systemen, doelgroepen en organisatorische volwassenheid.

Maak daarom vanaf de start helder wat centraal wordt vastgesteld, waar ministeries zelf invulling aan geven, wie waarvoor verantwoordelijk is en hoe afwijkingen worden beoordeeld. Zo ontstaat een gezamenlijke basis met voldoende flexibiliteit voor effectieve implementatie per ministerie.

Daarmee verschuift de opgave van een technologisch vraagstuk naar een ontwerpvraagstuk voor een duurzaam socio-technisch systeem. De volgende stap ligt daarmee niet in verdere analyse, maar in het expliciet maken van keuzes en het gecontroleerd toepassen in de praktijk, zodat de haalbaarheid onder realistische omstandigheden kan worden vastgesteld.

Bijlage

Dimensie	NA 2018	RDDI 2025	AZ 2025	RDDI 2026
Fase/ volwassenheid	Proof-of-concept	Verkennend (lab)	Uitwerking randvoorwaarden	Praktijkpilot
Technische status	Aangetoond (beperkt)	Werkend in experiment	Werkend onder voorwaarden	Bewezen in praktijk
Toepasbaarheid	Beperkt	Nog niet representatief	Voorwaardelijk toepasbaar	Selectief toepasbaar
Classificatie-complexiteit	Laag (binair)	Middel	Hoog (incl. bewaartermijnen)	Hoog (praktijkvariatie)
Model-afhankelijkheid	Hoog	Hoog	Relatief	Laag (data dominant)
Datakwaliteit	Beperkt en experimenteel	Niet representatief	Kritische succesfactor	Doorslaggevend
Prestatie-betrouwbaar-heid	Onzeker	Onzeker	Redelijk (taakafhankelijk)	Goed (voor grote categorieën)
Beperkingen (technisch)	Schaalbaarheid	Generaliseer-baarheid	Complexiteit	Kleine categorieën
Rol van de mens	Training essentieel	Ondersteunend	Controle noodzakelijk	Structureel onderdeel
Mate van automatisering	Laag	Ondersteunend	Gedeeltelijk	Selectief en risicogestuurd
Uitlegbaarheid/ vertrouwen	Aandachtspunt	Onderontwikkeld	Sterk uitgewerkt	Voorwaarde
Juridische haalbaarheid	Onzeker	Vermeden (synthetische data)	Uitgewerkt en mogelijk	Toepasbaar mits ingericht
Besluitvorming	Ondersteunend	Ondersteunend	Probabilistisch	Drempelwaarde + controle
Organisatorische impact	Nog onbekend	Beperkt inzicht	Procesverandering nodig	Complex en multidisciplinair
Implementatie-status	Niet aan de orde	Conceptueel	Ontwerpfase	Praktisch mogelijk
Schaalbaarheid	Onzeker	Onbewezen	Voorwaardelijk	Afhankelijk van data & governance
Kernbeperking	Vertrouwen	Data	Complexiteit	Data + organisatie
Risicoprofiel	Hoog (onzekerheid + geen controle-mechanismen)	Hoog (niet representatieve data)	Middel (afhanke-lijk van inrichting)	Beheersbaar (mits rand-voorwaarden)
Geschikt voor inzet?	Nee	Nog niet	Beperkt (pilot/ gecontroleerde context)	Ja, selectief en gefaseerd

Deze brochure is een uitgave van:

Rijksprogramma voor Duurzaam Digitale Informatiehuishouding (RDDI)

Oktober 2026