

BURGERVERTROUWEN IN PUBLIEKE ORGANISATIES NA EEN AI-FALEN

HET MODERERENDE EFFECT VAN HET TYPE FALEN OP DE
RELATIE TUSSEN DE ATTRIBUTIE IN EEN VERONTSCHULDIGING
EN VERTROUWENSHERSTEL

Aantal woorden: 14657

Alex Depinois

Studentennummer: 02111702

Promotor: Prof. dr. Tessa Haesevoets

Masterproef voorgelegd voor het behalen van de graad master in de Bestuurskunde en het Publiek Management

Academiejaar: 2025 – 2026

Voorwoord

Mijn interesse in behavioral public administration ligt aan de basis van dit werkstuk. Het vakgebied vertrekt van een eenvoudige maar verstrekkende vaststelling: bestuur is geen abstract systeem van regels en procedures, maar een aaneenschakeling van beslissingen en percepties van mensen. Wat een overheidsorganisatie op papier doet en wat een burger daarvan waarneemt, zijn twee verschillende dingen, en het is precies in die kloof dat vertrouwen wordt gewonnen of verloren. Dat inzicht toepassen op de opkomst van AI in de publieke dienstverlening leek mij niet alleen intellectueel prikkelend, maar ook urgent. Naarmate overheden beslissingen aan algoritmen toevertrouwen, wordt de vraag wie de verantwoordelijkheid draagt wanneer die beslissingen foutlopen minder vanzelfsprekend te beantwoorden. De experimentele methode van dit vakgebied bood mij de kans om die vraag niet enkel te beschrijven, maar effectief te toetsen.

Het schrijven van deze thesis was niet mogelijk geweest zonder een aantal belangrijke mensen, deze zet ik in dit voorwoord ook graag in de kijker. In het bijzonder wil ik mijn promotor, prof. dr. Tessa Haesevoets, bedanken. Haar eerder onderzoek naar vertrouwensherstel en naar de perceptie van AI in overheidsbeslissingen liet me toe haar feedback nauwgezet en met vertrouwen toe te passen. Daarnaast gaf ze me de sturing die ik nodig had om het onderzoek te concretiseren en aan te vatten, maar ook bij het uitwerken en finaliseren kon ik steeds rekenen op haar expertise.

Daarnaast gaat mijn dank uit naar de respondenten die de tijd namen om de vragenlijst in te vullen. Zonder hun bereidwilligheid was er geen mogelijkheid tot het volmaken van deze thesis.

Tot slot vermeld ik ook graag enkele mensen uit mijn persoonlijke kring, nu ik aangekomen ben bij het einde van deze studie kan ik stellen dat mijn ouders telkens mijn belangrijkste motivatie waren en ook bij het schrijven van deze thesis was hun hulp onmisbaar. Zo kon ik op hen rekenen voor het massaal verspreiden van de enquête en het ronselen van respondenten, alsook het nalezen en corrigeren van kleine spellingsfouten en andere details. Dit droeg uiteindelijk bij aan de kwaliteit van deze thesis, waarvoor heel veel dank.

Daarnaast wil ik graag mijn opa en oma bedanken voor hun onvoorwaardelijke steun en het zorgen voor een overvloed aan leesmateriaal omtrent mijn onderwerp. Zonder hen zou ik de moed niet gehad hebben om deze uitdaging aan te vatten. Diezelfde steun kreeg ik ook van mijn vriendin waar ik stevast naartoe ging wanneer ik het bos niet langer door de bomen zag, zij zorgde ervoor dat ik steeds in een rustgevende en liefdevolle omgeving kon schrijven.

De auteur en de promotor geven de toelating deze masterproef voor consultatie beschikbaar te stellen en delen van de masterproef te kopiëren voor persoonlijk gebruik. Elk ander gebruik valt onder de beperkingen van het auteursrecht, in het bijzonder met betrekking tot de verplichting de bron uitdrukkelijk te vermelden bij het aanhalen van resultaten uit de masterproef.

Ik verklaar dat het onderzoek verlopen is volgens de regels die gelden inzake wetenschappelijke en academische integriteit. Ik verklaar kennis te hebben genomen van, en gehandeld te hebben in overeenstemming met, de ethische code van de faculteit.

Inhoud

Voorwoord	2
1. Inleiding	6
1.1 Probleemstelling en maatschappelijke context.....	6
1.2 Wetenschappelijke relevantie en onderzoeksgap	7
1.3 Centrale onderzoeksvraag en deelvragen.....	8
1.4 Structuur van de masterproef	9
2. Literatuuroverzicht	10
2.1 Burgervertrouwen in publieke organisaties	10
2.1.1 Conceptualisering en dimensies van vertrouwen	10
2.1.2 Antecedenten van burgervertrouwen	11
2.1.3 Vertrouwen in de organisatie versus vertrouwen in de leider	11
2.2 Artificiële intelligentie in de publieke sector	12
2.2.1 De opmars van AI in publieke dienstverlening.....	12
2.2.2 Specifieke risico's van AI-systemen in publieke contexten.....	12
2.2.3 Empirische voorbeelden van AI-falen in publieke organisaties	13
2.3 Vertrouwensbreuk: types van organisationeel falen.....	14
2.3.1 De competentie-integriteit dichotomie.....	14
2.3.2 Asymmetrische gevolgen voor vertrouwen	14
2.3.3 Toepassing op AI-falen	14
2.4 Attributietheorie als theoretisch kader	15
2.4.1 Weiner's attributiemodel.....	15
2.4.2 Attributies bij organisationeel falen	15
2.4.3 Attributies en AI: de verschuiving van verantwoordelijkheid	16
2.5 Vertrouwensherstel en verontschuldigungsstrategieën	16
2.5.1 Typologie van herstelstrategieën	16
2.5.2 Interne versus externe attributie in verontschuldiging	17
2.5.3 Het interactie-effect van Kim et al. (2004, 2006)	17
2.5.4 Verontschuldiging na AI-falen: opkomende inzichten	18
2.6 Conceptueel kader en hypothesen	19
3. Methodologie	20
3.1 Onderzoeksstrategie en -ontwerp.....	20
3.2 Opbouw van het vignetmateriaal	20
3.3 Steekproef en dataverzameling	21
3.4 Operationalisering en meetinstrumenten.....	22
3.4.1 Afhankelijke variabele: burgervertrouwen.....	22

3.4.2 Onafhankelijke variabelen en manipulatiechecks	22
3.4.3 Controlevariabelen	22
3.5 Analysestrategie	22
4. Resultaten	24
4.1 Betrouwbaarheid van de meetinstrumenten	24
4.2 Steekproefbeschrijving en randomisatiecontrole	24
4.3 Beschrijvende statistieken en binnen-conditie veranderingen	25
4.4 Toetsing van de hypothesen: vertrouwen in de organisatie	29
4.5 Toetsing van de hypothesen: vertrouwen in de leidinggevende	30
4.6 Controle van de statistische aannames.....	31
4.7 Robuustheidsanalyse met covariaten.....	31
4.8 Samenvattend overzicht van de hypothesetoetsing	32
5. Discussie en conclusie	33
5.1 Interpretatie van de belangrijkste bevindingen	33
5.2 Vergelijking met de bestaande literatuur	34
5.3 Theoretische bijdrage	35
5.4 Praktische implicaties	36
5.5 Beperkingen van het onderzoek	37
5.6 Aanbevelingen voor toekomstig onderzoek.....	38
5.7 Conclusie	39
Literatuurlijst	41

1. Inleiding

1.1 Probleemstelling en maatschappelijke context

Vertrouwen in publieke instellingen vormt een fundamentele bouwsteen van democratische legitimiteit en effectieve beleidsuitvoering. Wanneer burgers hun overheid vertrouwen, zijn zij meer geneigd om wetten na te leven, belastingen correct af te dragen en mee te werken aan collectieve voorzieningen (Van de Walle & Bouckaert, 2003). Het laatste decennium tonen surveys zoals het OECD *Trust in Government*-rapport en de Edelman Trust Barometer dat dit vertrouwen in vele Westerse democratieën onder druk staat, met regelmatige dips naar aanleiding van politieke schandalen, beleidsmislukkingen of mediagenieke incidenten (OECD, 2022). Voor België specifiek wijzen Van de Walle en Bouckaert (2003) op de paradox dat objectieve overheidsprestaties slechts beperkt voorspellen hoe burgers hun overheid percipiëren. Perceptie wordt minstens evenzeer gevormd door media-aandacht, schandalen en de manier waarop bestuurders communiceren na incidenten.

De laatste vijf jaar heeft een nieuwe ontwikkeling deze klassieke vertrouwensdynamiek fundamenteel veranderd: de grootschalige adoptie van kunstmatige intelligentie (AI) en geautomatiseerde besluitvorming door publieke organisaties. Overheden zetten AI in voor uiteenlopende taken, gaande van fraudedetectie in sociale zekerheid en belastingen, over geautomatiseerde toekenning van uitkeringen en visa, tot voorspellende systemen in politiezorg en jeugdhulp (Wirtz et al., 2019; Young et al., 2019). De verwachte voordelen, namelijk efficiëntie, consistentie en het wegnemen van menselijke biases, worden zelden zonder kanttekeningen gerealiseerd. AI-systemen falen op manieren die fundamenteel verschillen van traditioneel menselijk falen: zij produceren systematische fouten op grote schaal, zijn vaak ondoorzichtig voor zowel burgers als ambtenaren, en raken vaak disproportioneel de meest kwetsbare bevolkingsgroepen (Alon-Barkat & Busuioc, 2023; Busuioc, 2021).

Het bekendste voorbeeld van de afgelopen jaren is het Nederlandse toeslagenschandaal, waarbij een zelflerend risicomodel van de Belastingdienst gedurende jaren tienduizenden ouders ten onrechte als fraudeur bestempelde. Variabelen zoals dubbele nationaliteit fungeerden als impliciete risico-indicatoren, met als gevolg dat vooral gezinnen met een migratieachtergrond werden getroffen (Amnesty International, 2021; van Bakkum & Zuiderveen Borgesius, 2021). Het schandaal leidde in januari 2021 tot het aftreden van de regering-Rutte III. In dezelfde periode oordeelde de Haagse rechtbank in een baanbrekend vonnis dat het Nederlandse profileringssysteem SyRI in strijd was met artikel 8 EVRM (van Bakkum & Zuiderveen Borgesius, 2021). Ook buiten Nederland zijn vergelijkbare schandalen gedocumenteerd: het Australische *Robodebt*-programma beschuldigde tussen 2015 en 2019 ongeveer 500.000 uitkeringsgerechtigden onterecht van fraude, en in Vlaanderen rezen vergelijkbare bezorgdheden rond de inzet van profileringssytemen door de VDAB en de Kruispuntbank van de Sociale Zekerheid (Selten et al., 2023).

Deze incidenten zijn niet louter technische ongelukken. Zij raken de kernwaarden waarvoor publieke organisaties geacht worden te staan: gelijke behandeling, rechtvaardigheid, transparantie en menselijke waardigheid. De Amnesty-rapportage over het Nederlandse toeslagenschandaal stelt expliciet dat het black-boxkarakter van het algoritme leidde tot een “black hole of accountability”, waarbij niemand verantwoordelijkheid wilde of kon nemen (Amnesty International, 2021). Het is precies in deze nasleep dat het vraagstuk van vertrouwensherstel acuut wordt: hoe communiceren publieke organisaties en hun leiders nadat AI-systemen burgers schade hebben berokkend? Welke verontschuldigungsstrategie genereert het meeste vertrouwensherstel? En verschilt dit naargelang het type fout dat het AI-systeem maakte?

De wijze waarop een organisatie zich verontschuldigt en hoe zij de oorzaak van het falen attribueert, is geen detail. Sinds het werk van Kim et al. (2004) is bekend dat verschillende herstelstrategieën anders uitpakken naargelang het type vertrouwensbreuk: een verontschuldiging herstelde vertrouwen effectiever na een competentiefalen, terwijl een ontkenning effectiever was na een integriteitsfalen. In een vervolgstudie toonden Kim et al. (2006) aan dat ook de attributie binnen een verontschuldiging ertoe doet: bij een competentiefalen (“we konden het niet”) herstelde een verontschuldiging met interne attributie meer vertrouwen, terwijl bij een integriteitsfalen (“we wilden het verkeerd doen”) een verontschuldiging met externe attributie effectiever bleek. Dit interactie-effect is uitvoerig getest in private, interpersoonlijke en human-resourcecontexten, maar systematisch onderzoek naar dezelfde dynamiek in een publieke sector waarin AI als oorzakelijke factor figureert, ontbreekt grotendeels. Net daar wil deze masterproef een bijdrage leveren.

1.2 Wetenschappelijke relevantie en onderzoeksgap

De wetenschappelijke literatuur over institutioneel vertrouwen en vertrouwensherstel kent drie grotendeels onafhankelijk ontwikkelende stromingen. Een eerste stroming, gevoed door Mayer et al. (1995) en latere bestuurskundige toepassingen (Grimmelikhuijsen & Knies, 2017; Grimmelikhuijsen et al., 2013), onderzoekt de antecedenten en dimensies van burgervertrouwen in publieke organisaties. Een tweede stroming, vertrekkend van Lewicki en Bunker (1996), Kim et al. (2004, 2006) en Dirks et al. (2009), focust op vertrouwensbreuken en herstelstrategieën, voornamelijk in interpersoonlijke en organisationele settings. Een derde stroming, die de laatste vijf jaar in een stroomversnelling kwam, bestudeert burgerpercepties van AI-systemen in publieke dienstverlening (Alon-Barkat & Busuioc, 2023; Aoki, 2020; Grimmelikhuijsen, 2023; Schiff et al., 2022).

De integratie van deze drie stromingen is echter beperkt. De klassieke vertrouwensherstelliteratuur ontstond in een tijd waarin de vraag “wie veroorzaakte de fout?” meestal een eenduidig menselijk antwoord had: een werknemer, een leidinggevende, een organisatie als geheel. In het AI-tijdperk verandert dit. Wanneer een algoritme een verkeerde beslissing neemt, vermenigvuldigen de potentiële schuldigen zich: de ontwikkelaars, de bestellende overheid, de leidinggevende ambtenaar die het systeem inzette, de softwareleverancier, of zelfs “het algoritme” als een quasi-autonoom subject (Busuioc, 2021; de Bruijn et al., 2022). Hierdoor wordt attributie zowel cognitief complexer voor burgers als strategisch gevoeliger voor

publieke leiders. Een verontschuldiging waarin de leider stelt “het was onze fout” (interne attributie) impliceert dat de organisatie de keuze om AI in te zetten zelf draagt. Een verontschuldiging die “de algoritmes”, “het systeem” of “de dataleverancier” in een externe positie plaatst, kan ontvangen worden als verantwoordelijkheidsontwijking.

Beperkt empirisch onderzoek heeft de combinatie van attributietheorie en AI-falen bestudeerd, voornamelijk in commerciële of interpersoonlijke contexten (Kim & Song, 2021; Glikson & Woolley, 2020). Kim en Song (2021) toonden bijvoorbeeld dat interne attributies effectiever zijn bij sterk antropomorfe AI-agenten, terwijl externe attributies beter werken bij minder mensachtige systemen. Deze inzichten zijn echter moeilijk één-op-één over te dragen naar de publieke sector, waar de relatie tussen burger en organisatie fundamenteel anders is dan tussen consument en bedrijf. Burgers hebben geen “exit”-optie tegenover hun overheid; zij zijn afhankelijk van publieke organisaties voor uitkeringen, identiteitsdocumenten, gezondheidszorg en justitie. Bovendien gelden voor de overheid hogere morele en juridische standaarden van non-discriminatie, transparantie en rechtszekerheid (Meijer & Grimmelikhuijsen, 2020). Dit suggereert dat de manier waarop burgers verontschuldigingen evalueren in het licht van AI-falen, niet zomaar voorspeld kan worden vanuit private-sectoronderzoek.

De onderzoeksgap die deze masterproef invult, situeert zich dus op het kruispunt van drie domeinen: (1) de bestuurskundige literatuur over burgervertrouwen, (2) de psychologische en managementliteratuur over vertrouwensherstel en attributietheorie, en (3) de opkomende literatuur over AI in het openbaar bestuur. Concreet ontbreekt experimenteel bewijs over de vraag of het klassieke interactie-effect van Kim et al. (2006), interne attributie effectiever bij competentiefalen, externe bij integriteitsfalen, ook van toepassing is wanneer (a) de organisatie publiek is, (b) het falen een AI-systeem betreft, en (c) zowel het vertrouwen in de organisatie als in de individuele leider gemeten wordt. Bovendien biedt het Belgische empirische landschap een waardevolle context: relatief weinig studies hebben Belgische respondenten gebruikt om hypothesen over algoritmische besluitvorming te testen, terwijl de Vlaamse en federale overheden in toenemende mate AI inzetten (Selden et al., 2023; Wirtz et al., 2019).

1.3 Centrale onderzoeksvraag en deelvragen

Op basis van de geschetste maatschappelijke en wetenschappelijke relevantie wordt in deze masterproef de volgende centrale onderzoeksvraag geformuleerd:

“In welke mate heeft het type falen (competentie- versus integriteitsfout) van een AI-systeem in een publieke organisatie een modererend effect op de relatie tussen de attributie in een verontschuldiging (interne versus externe attributie) en het herstel van burgervertrouwen in de organisatie en haar leider?”

Deze centrale onderzoeksvraag wordt verder opgesplitst in drie deelvragen:

- DV1.** In welke mate leidt een AI-gerelateerd integriteitsfalen tot een sterkere daling van het burgersvertrouwen in een publieke organisatie en haar leider dan een AI-gerelateerd competentiefalen?
- DV2.** Is een verontschuldiging met interne attributie effectiever dan één met externe attributie voor het herstellen van burgersvertrouwen na een AI-falen in een publieke organisatie?
- DV3.** Verschilt het effect van de attributie in een verontschuldiging op vertrouwensherstel naargelang het type falen (competentie versus integriteit) van het AI-systeem?

Het onderzoek wordt operationeel ingevuld via een online vignetonderzoek met een 2 (type falen: competentie vs. integriteit) × 2 (attributie: intern vs. extern) between-subjects design, waarbij Belgische burgers (18+) ad random aan één van vier condities worden toegewezen. Dit experimenteel design sluit nauw aan bij de methodologische traditie van Kim et al. (2004) en bij recente vignetstudies in de publieke administratie (Alon-Barkat & Busuioc, 2023; Grimmelikhuijsen, 2023). De methodologische operationalisering wordt uitgebreid besproken in Hoofdstuk 3.

1.4 Structuur van de masterproef

Deze masterproef is als volgt opgebouwd. Hoofdstuk 2 biedt een uitgebreid literatuuroverzicht waarin volgende zaken worden besproken: het concept burgersvertrouwen en zijn dimensies (2.1), de opkomst van AI in publieke dienstverlening en de specifieke risico's die hiermee gepaard gaan (2.2), de typologie van organisationeel falen (2.3), attributietheorie als verklarend kader (2.4) en de literatuur over vertrouwensherstel en verontschuldigungsstrategieën (2.5). Het hoofdstuk sluit af met het conceptueel kader en de geformuleerde hypothesen (2.6). Hoofdstuk 3 beschrijft de methodologie, waaronder het experimentele design, de vignetmaterialen, de meetinstrumenten en de gehanteerde analysetechnieken. Hoofdstuk 4 rapporteert de empirische resultaten, en Hoofdstuk 5 bevat de discussie, theoretische bijdrage, praktische implicaties, beperkingen en aanbevelingen voor toekomstig onderzoek.

2. Literatuuroverzicht

Dit hoofdstuk schetst het theoretisch kader waarbinnen de onderzoeksvraag wordt gesitueerd. De opbouw is als volgt: eerst wordt het centrale concept burgersvertrouwen gedefinieerd en geoperationaliseerd (2.1), vervolgens wordt de specifieke context van AI in de publieke sector geïntroduceerd, inclusief empirische voorbeelden van AI-falen (2.2). Daarna wordt de literatuur over organisationeel falen besproken (2.3) en wordt attributietheorie geïntroduceerd als verklarend mechanisme (2.4). Sectie 2.5 integreert deze inzichten in een bespreking van vertrouwensherstelstrategieën, in het bijzonder verontschuldigen, en behandelt de specifieke uitdagingen van vertrouwensherstel na AI-falen. Sectie 2.6 vat alles samen in een conceptueel model en formuleert drie hypothesen.

2.1 Burgersvertrouwen in publieke organisaties

2.1.1 Conceptualisering en dimensies van vertrouwen

Vertrouwen wordt in de organisatiepsychologische literatuur klassiek gedefinieerd als “the willingness of a party to be vulnerable to the actions of another party based on the expectation that the other will perform a particular action important to the trustor, irrespective of the ability to monitor or control that other party” (Mayer et al., 1995, p. 712). Centraal staan dus twee elementen: een bereidheid tot kwetsbaarheid en een positieve verwachting ten aanzien van het toekomstig gedrag van de vertrouwde partij. Toegepast op de relatie tussen burger en publieke organisatie betekent dit dat een burger erop vertrouwt dat de overheid haar bevoegdheden gebruikt op een manier die het algemeen belang dient en de eigen belangen van de burger niet schaadt, zelfs wanneer de burger geen directe controle heeft over die uitvoering (Grimmelikhuijsen & Knies, 2017).

Mayer et al. (1995) onderscheiden drie kerndimensies van waargenomen betrouwbaarheid: vermogen (*ability*), welwillendheid (*benevolence*) en integriteit (*integrity*). Vermogen verwijst naar de competenties, vaardigheden en kenmerken waardoor de organisatie haar taken effectief kan uitvoeren. Welwillendheid betreft de mate waarin de organisatie het goed voor heeft met de burger en niet louter eigen of egoïstische belangen dient. Integriteit gaat over de naleving van morele principes die de burger als wenselijk beschouwt, zoals eerlijkheid, gelijke behandeling en rechtvaardigheid. Hoewel deze drie dimensies in empirisch onderzoek vaak hoog samenhangen, hebben zij conceptueel een onderscheiden basis: een ambtenaar kan bijvoorbeeld bekwaam zijn (hoog op *ability*), maar gepercipieerd worden als bevooroordeeld (laag op *integrity*).

Grimmelikhuijsen en Knies (2017) ontwikkelden en valideerden een schaal voor burgersvertrouwen in publieke organisaties die expliciet voortbouwt op het model van Mayer et al. (1995). Hun werk toont aan dat het drie-dimensionele model robuust is in zowel Westerse als niet-Westerse contexten en bruikbaar is voor vergelijkend bestuurskundig onderzoek. In een gerelateerd cross-cultureel experiment lieten Grimmelikhuijsen et al. (2013) zien dat transparantie het vertrouwen in de overheid versterkt, maar dat het

effect afhankelijk is van pre-existerend cynisme van de respondent: meer transparantie kan bij sceptische burgers paradoxaal leiden tot minder vertrouwen, indien de blootgelegde informatie problematisch is. Dit nuanceert de naïeve assumptie dat “meer openheid” automatisch “meer vertrouwen” genereert en heeft directe gevolgen invloed op communicatie van publieke organisaties na een incident.

2.1.2 Antecedenten van burgersvertrouwen

Het bestuurskundig onderzoek heeft een veelheid aan antecedenten van burgersvertrouwen geïdentificeerd. Klassieke antecedenten omvatten prestatie (de mate waarin de organisatie haar taken effectief uitvoert), procedurele rechtvaardigheid (de mate waarin processen als eerlijk worden ervaren), responsiviteit (de mate waarin de organisatie reageert op burgers) en transparantie (de mate waarin de organisatie inzicht biedt in haar werking) (Grimmelikhuijsen et al., 2013; Van de Walle & Bouckaert, 2003). Van de Walle en Bouckaert (2003) beklemtoonden dat objectieve prestaties slechts een deel van de variantie in waargenomen vertrouwen verklaren, daarnaast spelen ook macro-economische factoren, mediarapportering en gepolitiseerde culturele scripts een rol.

In de afgelopen jaren is daar een nieuwe antecedent bijgekomen: de waargenomen kwaliteit en eerlijkheid van algoritmische en geautomatiseerde besluitvorming (Grimmelikhuijsen, 2023; Schiff et al., 2022). Onderzoek van Grimmelikhuijsen (2023) toont aan dat explainability van algoritmes een sterkere invloed heeft op vertrouwen dan toegankelijkheid van broncode of documentatie. Burgers willen weten waarom een beslissing is genomen, niet zozeer hoe het systeem technisch werkt. Schiff et al. (2022) tonen daarnaast aan dat “transparency failure”, situaties waarin de bevoegde ambtenaren de algoritmische beslissing niet kunnen verklaren, zelfs tot een bijzonder scherpe daling van vertrouwen leidt. Dit suggereert dat de verwachtingen van burgersvertrouwen voor AI-gemedieerde dienstverlening verbreden, met explainability en accountability als nieuwe, kritische dimensies.

2.1.3 Vertrouwen in de organisatie versus vertrouwen in de leider

Een belangrijk onderscheid dat in deze masterproef expliciet gemaakt wordt, is dat tussen vertrouwen in de organisatie als entiteit en vertrouwen in de individuele leider (CEO, minister, dienstverantwoordelijke). Hoewel beide vormen van vertrouwen vaak correleren, kunnen zij divergeren in reactie op een incident (Dirks et al., 2009). Een organisatie kan beschadigd geraken in haar institutionele reputatie, terwijl een sterke en transparant communicerende leider net vertrouwenswinst boekt. Omgekeerd kan een geloofwaardige organisatie afbrokkelen wanneer haar leider met een persoonlijk schandaal in verband wordt gebracht. Dit onderscheid is voor de huidige studie van bijzonder belang omdat verontschuldiging na AI-falen typisch zowel door de organisatie als door één of meer leiders worden uitgesproken, en de attributiekeuze (“ik”, “wij”, “het algoritme”, “de leverancier”) elk van die twee vertrouwenslagen anders kan raken.

2.2 Artificiële intelligentie in de publieke sector

2.2.1 De opmars van AI in publieke dienstverlening

De inzet van AI door publieke organisaties is in het laatste decennium exponentieel gegroeid. Wirtz et al. (2019) onderscheiden vier toepassingsdomeinen: (1) interne automatisering van routinetaken, (2) ondersteuning van beleidsbeslissingen via voorspellende analytics, (3) directe burgerinteractie via chatbots en virtuele assistenten, en (4) toezicht en handhaving, bijvoorbeeld bij fraudedetectie. Young et al. (2019) introduceren in dit verband het concept “artificial discretion”, waarbij beslissingsbevoegdheid die voorheen bij ambtenaren lag, gedeeltelijk wordt overgedragen aan algoritmische systemen. Dit is een fundamentele verschuiving van wat Bovens en Zouridis (2002) ruim twintig jaar geleden al voorspelden als de overgang van “street-level” naar “system-level” bureaucratie.

Deze opmars wordt gedreven door een mix van factoren: budgettaire druk op overheden, beloftes van efficiëntiewinst door techbedrijven, het bestaan van grote administratieve datasets, en politiek populaire frames van “fraudebestrijding” en “modernisering” (Busuioc, 2021; Janssen & van den Hoven, 2015). Tegelijk staat het wetgevend kader nog in opbouw. De Europese AI Act, definitief aangenomen in 2024, categoriseert AI-toepassingen in publieke dienstverlening; met name in domeinen zoals migratie, justitie en sociale zekerheid als hoog-risico. Daarnaast legt het verplichtingen op rond transparantie, menselijke supervisie en risicobeoordeling (Europese Commissie, 2024). Voor leiders van publieke organisaties betekent dit dat de governance van AI niet meer optioneel is, maar een kerntaak wordt.

2.2.2 Specifieke risico's van AI-systemen in publieke contexten

Hoewel AI verschillende voordelen kan bieden, identificeert de literatuur een aantal specifieke risico's die particulier zijn voor de publieke sector (Busuioc, 2021; Meijer & Grimmelhuijsen, 2020). Een eerste risico is bias en discriminatie: AI-systemen trainen op historische data, en wanneer die data systematische ongelijkheden bevatten (bijvoorbeeld een overrepresentatie van bepaalde etnische groepen in fraudedossiers), reproduceren en versterken zij die patronen. Het Nederlandse toeslagenschandaal is hiervan een paradigmatisch voorbeeld, waarbij dubbele nationaliteit als risicofactor fungeerde en de toepassing dus systematisch personen met een migratieachtergrond raakte (Amnesty International, 2021).

Een tweede risico betreft accountability. In klassieke administratieve beslissingen kan een gedupeerde burger de redenering van de ambtenaar laten toetsen, motivering eisen en in beroep gaan. Bij ondoorzichtige machine-learningmodellen is dit veel moeilijker (Busuioc, 2021; de Bruijn et al., 2022). De Bruijn et al. (2022) wijzen er bovendien op dat ook “explainable AI” niet zonder paradoxen is. Uitleg kan strategisch gegeven worden om controverses te dempen zonder realistische controlemogelijkheid, of kan de burger overbelasten met technische details die hem opnieuw geen handelingsperspectief bieden.

Een derde risico is automation bias. Dit is het verschijnsel waarbij menselijke beslissers algoritmische adviezen onvoldoende kritisch beoordelen, ook wanneer er duidelijke signalen zijn dat het algoritme fout zit (Alon-Barkat & Busuioc, 2023). Alon-Barkat en Busuioc (2023) tonen experimenteel aan dat ook “selective adherence” optreedt: ambtenaren volgen algoritmische adviezen vooral wanneer die hun pre-existente stereotypen bevestigen, met als gevolg dat AI bestaande vooroordelen kan legitimeren onder de schijn van objectiviteit.

Samen verschuiven deze risico's de aard van organisationeel falen. Waar traditioneel falen vaak lokaal en zichtbaar is (“deze ambtenaar maakte een fout in dit dossier”), is AI-falen systemisch, ondoorzichtig en gemakkelijk te ontkennen of door te schuiven naar technische of externe actoren. Dit heeft directe implicaties voor de attributievraag die centraal staat in deze masterproef.

2.2.3 Empirische voorbeelden van AI-falen in publieke organisaties

Naast het reeds besproken Nederlandse toeslagenschandaal en het SyRI-systeem zijn de afgelopen jaren wereldwijd meerdere casussen van AI-falen in publieke contexten gedocumenteerd. Het Australische Robodebt-programma genereerde tussen 2015 en 2019 honderdduizenden incorrecte schulden bij uitkeringsgerechtigden en leidde tot een nationale Royal Commission die scherpe kritiek formuleerde op de bestuurlijke verantwoordelijkheid (van Bekkum & Zuiderveen Borgesius, 2021). In de Verenigde Staten leverde het COMPAS-systeem voor recidiveprognoses systematisch hogere risicoscores op voor zwarte verdachten, ook wanneer hun feitelijke recidive niet hoger lag (Busuioc, 2021). In het Verenigd Koninkrijk leidde het zogenaamde “A-level algorithm” van 2020, dat examenscores van middelbare scholieren voorspelde tijdens de coronacrisis, tot massale woede toen bleek dat het systematisch leerlingen uit kansarme scholen benadeelde (Selten et al., 2023).

In de Vlaamse en Belgische context is publiek empirisch onderzoek schaarser, maar bezorgdheden over algoritmische profileringssystemen bij onder meer de VDAB, de Kruispuntbank van de Sociale Zekerheid en lokale OCMW's zijn gedocumenteerd door zowel academici als belangenorganisaties. Het ontbreken van een grootschalig Belgisch “toeslagenschandaal” sluit allerminst uit dat vergelijkbare risico's sluimeren. Het maakt de wetenschappelijke vraag hoe Belgische burgers reageren op AI-falen door hun overheid des te relevanter, omdat hun reacties nog niet door één dominante mediagebeurtenis vastgelegd zijn. Wat deze casussen verbindt, is dat zij allen vragen oproepen die voor deze masterproef centraal staan: ging het om een competentiefalen (het systeem werkte technisch niet zoals beoogd) of om een integriteitsfalen (het systeem werd ingezet op een manier die discriminatie of disproportionele schade in de hand werkte)? En hoe communiceerden de betrokken leiders nadien?

2.3 Vertrouwensbreuk: types van organisationeel falen

2.3.1 De competentie-integriteit dichotomie

Een centrale theoretische bijdrage van Kim et al. (2004) is het onderscheid tussen *competence-based* en *integrity-based* vertrouwensbreuken. Een competentiefalen ontstaat wanneer de trustee niet over de capaciteiten beschikte om de gewenste handeling uit te voeren: een arts die door onkunde een verkeerde diagnose stelt, een aannemer die door technisch onvermogen een gebouw verkeerd construeert, een algoritme dat door slechte training systematisch verkeerde voorspellingen produceert. Een integriteitsfalen ontstaat daarentegen wanneer de trustee de capaciteit wel had, maar bewust of door ethisch onverantwoorde keuzes morele principes schond: een arts die een verkeerde diagnose verzint om financieel gewin, een aannemer die spaart op veiligheid, een overheidsorganisatie die bewust een discriminerend algoritme inzet.

Dit onderscheid sluit conceptueel aan bij twee van de drie dimensies van Mayer et al. (1995): competentiefalen tast vooral de *ability*-dimensie aan, integriteitsfalen vooral de *integrity*-dimensie. De welwillendheidsdimensie (*benevolence*) kan door beide getroffen worden, maar wordt typisch sterker geraakt door integriteitsfalen. Empirisch blijkt dat hetzelfde gedrag (bv. een foutief ingediende belastingaangifte) naargelang de framing consistent als een competentie- dan wel een integriteitsfalen wordt herkend, wat de validiteit van het onderscheid ondersteunt (Kim et al., 2004, 2006).

2.3.2 Asymmetrische gevolgen voor vertrouwen

Een belangrijke bevinding in de literatuur is dat competentie- en integriteitsfouten asymmetrische gevolgen hebben voor vertrouwen. Integriteitsfouten leiden doorgaans tot een sterkere en duurzamere daling van vertrouwen dan competentiefouten van vergelijkbare omvang (Dirks et al., 2009; Kim et al., 2004, 2006).

Een tweede asymmetrie betreft de aard van het herstelproces. Competentiefouten kunnen op termijn worden hersteld door bewijs van leervermogen, training of corrigerende investeringen: “we waren te zwak, maar we hebben ons verbeterd” is een geloofwaardig narratief. Integriteitsfouten zijn moeilijker te “herstellen” omdat zij een uitspraak doen over het morele karakter van de organisatie. Pas door langdurige consistente ethische praktijk en symbolische gestes (bekentenis, restitutie, structurele hervorming) kan dit vertrouwen worden teruggewonnen (Dirks et al., 2009; Tomlinson et al., 2004). Voor leiders van publieke organisaties is dit verschil cruciaal omdat het bepaalt welke communicatiestrategie als geloofwaardig kan worden gepercipieerd.

2.3.3 Toepassing op AI-falen

Wanneer we de competentie-integriteit dichotomie toepassen op AI-falen, ontstaat een specifieke dynamiek. Veel AI-incidenten kunnen, afhankelijk van framing, als beide types worden voorgesteld. Het Nederlandse toeslagenschandaal kan beschouwd worden als een competentiefalen (“het model werd niet goed getest”), maar evengoed als een integriteitsfalen (“men wist of had moeten weten dat dubbele nationaliteit als variabele discriminerend werkt en koos toch om door te gaan”) (Amnesty International, 2021; van Bakkum &

Zuiderveen Borgesius, 2021). Deze framingflexibiliteit is precies wat communicatieadviseurs van publieke organisaties uitnodigt tot strategische keuzes: een leider die het incident als technisch competentieprobleem kadert, ontwijkt deels het zwaardere integriteitsverwijt. Een burger of journalist die het incident als integriteitsfalen kadert, dwingt daarentegen een veel zwaardere verantwoording af.

Vanuit het perspectief van het experiment in deze masterproef is het methodologisch cruciaal om de twee types scherp te manipuleren in de vignetmaterialen. Een competentievignet beschrijft een AI-systeem dat door technische fouten of onvolledige data systematisch verkeerde uitkomsten genereerde. Een integriteitsvignet beschrijft een AI-systeem dat ingezet werd op een wijze waarvan de organisatie redelijkerwijs kon weten dat het discriminerend, onethisch of disproportioneel was. De keuze van “een dienst sociale zaken” als context is hiervoor gepast, omdat deze sector zowel technisch-administratieve als ethisch-morele dimensies combineert.

2.4 Attributietheorie als theoretisch kader

2.4.1 Weiner's attributiemodel

De attributietheorie van Weiner (1986) biedt een fundamenteel kader voor het begrijpen van hoe individuen oorzaken toekennen aan gebeurtenissen, en hoe die attributies hun emotionele en gedragsmatige reacties beïnvloeden. Weiner onderscheidt drie centrale dimensies waarop attributies kunnen variëren: *locus of control* (intern versus extern: ligt de oorzaak bij de actor of bij omgevingsfactoren?), *stability* (stabil versus instabil: is de oorzaak een blijvend kenmerk of een tijdelijke omstandigheid?), en *controllability* (controleerbaar versus oncontroleerbaar: had de actor de oorzaak kunnen voorkomen?). Hoewel alle drie dimensies relevant zijn voor het bestuderen van organisationele communicatie na een falen, focust de huidige studie primair op de locus-dimensie, conform de meeste bestaande herstelliteratuur (Kim et al., 2006; Kim & Song, 2021).

De locus-dimensie heeft directe affectieve consequenties. Wanneer burgers de oorzaak van een falen intern aan de organisatie attribueren, worden gevoelens van boosheid en verontwaardiging gewekt, omdat de organisatie als verantwoordelijk wordt gezien voor een vermijdbare situatie. Wanneer zij de oorzaak extern attribueren, wordt sympathie opgewekt, omdat de organisatie als slachtoffer van omstandigheden wordt gezien (Weiner, 1986). Deze emoties bepalen vervolgens mede het bereidheidsniveau tot vertrouwensherstel.

2.4.2 Attributies bij organisationeel falen

Toegepast op organisationeel falen onderscheidt de literatuur tussen attributies die de organisatie zelf maakt (in haar communicatie) en attributies die observatoren spontaan formuleren (Coombs, 2007). De Situational Crisis Communication Theory van Coombs (2007) stelt dat de organisatie haar communicatie idealiter laat aansluiten bij de attributies die het publiek waarschijnlijk reeds maakt: bij “victim crises” (waar de organisatie zelf slachtoffer lijkt, bv. natuurramp) kan een verdedigende strategie volstaan; bij “preventable crises” (waar

de organisatie als verantwoordelijk wordt gezien, bv. negligence) is een berouwvolle strategie met interne attributie noodzakelijk.

Wanneer een publieke organisatie in een verontschuldiging een interne attributie gebruikt, erkent zij dat de oorzaak binnen haar eigen invloedssfeer lag. Dit signaleert verantwoordelijkheid en kan integriteit demonstreren, maar verhoogt tegelijk de schuldperceptie en kan competentievragen oproepen (Kim et al., 2004). Wanneer zij daarentegen een externe attributie hanteert, beschermt zij haar reputatie tegen verwijten van wangedrag of onkunde, maar riskeert zij als ontwijkend, defensief of niet-empathisch over te komen, vooral bij ernstige incidenten (Coombs, 2007; Kim & Song, 2021). De juiste keuze is dus context-afhankelijk, en het is precies die contextafhankelijkheid die het interactie-effect met het type falen suggereert.

2.4.3 Attributies en AI: de verschuiving van verantwoordelijkheid

De aanwezigheid van een AI-systeem in de keten van besluitvorming introduceert een nieuwe attributiekandidaat die in klassieke modellen ontbreekt. Recent onderzoek (Busuioc, 2021; de Bruijn et al., 2022; Glikson & Woolley, 2020) wijst op een fenomeen dat sommige auteurs “moral crumple zone” of “algorithmic scapegoating” noemen: de neiging om verantwoordelijkheid voor een onwenselijke uitkomst af te schuiven naar het algoritme zelf, dat in de discoursvorming functioneert als een quasi-autonoom, schuldig agent. Voor publieke leiders biedt dit een verleidelijke retorische uitweg: “niet wij, maar het algoritme”. De vraag is echter of burgers deze externe attributie als geloofwaardig en gerechtvaardigd accepteren, of net als verantwoordelijkheidsontvluchting evalueren.

Kim en Song (2021) toonden in een experimenteel onderzoek aan dat de aanvaardbaarheid van externe attributies door AI-agenten zelf afhangt van de mate van antropomorfisme. Bij sterk “mensachtige” AI-agenten leidt externe attributie tot wantrouwen, terwijl bij “machine-achtige” AI-agenten externe attributie net effectiever is. Deze bevinding suggereert dat de manier waarop een publieke organisatie haar AI-systeem voorstelt, als een neutraal technisch instrument of als een quasi-beleidsmaker, de aanvaardbaarheid van haar attributiestrategie beïnvloedt. In een publieke context, waar morele verwachtingen hoog liggen, ligt de hypothese voor de hand dat externe attributies sneller als problematisch worden gepercipieerd, vooral wanneer het falen ernstig en moreel gekleurd is.

2.5 Vertrouwensherstel en verontschuldigungsstrategieën

2.5.1 Typologie van herstelstrategieën

Lewicki en Bunker (1996) waren onder de eersten die systematisch beschreven hoe geschonden vertrouwen kan worden hersteld. Zij identificeerden vier fasen: (1) erkenning van de overtreding, (2) bepaling van de oorzaak, (3) acceptatie van verantwoordelijkheid, en (4) herstelactie. Latere literatuur heeft deze fasen geoperationaliseerd in concrete strategieën die organisaties kunnen inzetten: ontkenning (“er was geen fout”), verontschuldiging (“het spijt ons”), compensatie (“we vergoeden de schade”), uitleg (“hier is wat er

gebeurde”) en structurele hervorming (“we veranderen het systeem”) (Dirks et al., 2009; Tomlinson et al., 2004). Deze strategieën sluiten elkaar niet uit; in de praktijk gebruiken organisaties vaak combinaties.

Empirisch onderzoek toont dat verontschuldiging in de meeste contexten de meest effectieve eenvoudige strategie is voor het herstellen van vertrouwen, althans wanneer zij oprecht overkomt en gepaard gaat met blijken van leervermogen of structurele actie (Dirks et al., 2009). Ontkenning kan effectief zijn wanneer de organisatie aantoonbaar onschuldig is, maar slaat terug wanneer later bewijs van schuld opduikt. Compensatie helpt vooral wanneer de materiële schade duidelijk is, maar volstaat zelden alleen voor het herstel van emotionele en symbolische schade. Voor de masterproef focussen we op verontschuldiging als herstelstrategie, omdat deze in de praktijk de dominante respons is van publieke organisaties op AI-incidenten en omdat zij de attributiekeuze centraal stelt.

2.5.2 Interne versus externe attributie in verontschuldigingen

Een verontschuldiging is meer dan het uitspreken van de woorden “het spijt ons”. Effectieve verontschuldigingen bevatten typisch vier elementen: erkenning van het falen, expressie van spijt, belofte van beterschap, en cruciaal voor de huidige studie: een attributie van de oorzaak (Tomlinson et al., 2004). Het is in dat vierde element dat publieke leiders strategische keuzes maken die het effect van hun verontschuldiging fundamenteel bepalen.

Een interne attributie in een verontschuldiging signaleert dat de organisatie de fout zelf veroorzaakte. Een externe attributie schuift de oorzaak naar factoren buiten de organisatie. De twee strategieën hebben verschillende perceptuele consequenties. Interne attributie wordt door observatoren typisch als eerlijker en moreel sterker waargenomen, maar verhoogt ook de perceptie van schuld en, in sommige gevallen, van incompetentie of wangedrag (Kim et al., 2004, 2006). Externe attributie beschermt de competentie- en integriteitsreputatie, maar wordt sneller als ontwijkend, defensief of zelfs cynisch geïnterpreteerd, zeker wanneer slachtoffers reëel leed hebben ervaren (Coombs, 2007).

2.5.3 Het interactie-effect van Kim et al. (2004, 2006)

De kerntheoretische bijdrage van Kim et al. (2004) is dat het effect van een herstelstrategie afhangt van het type vertrouwensbreuk. In hun experimenten varieerden zij twee factoren: het type fout (competentie versus integriteit) en de gehanteerde herstelstrategie (verontschuldiging of ontkenning). Zij vonden dat vertrouwen effectiever werd hersteld met een verontschuldiging bij een competentiefalen, maar met een ontkenning bij een integriteitsfalen. De onderliggende logica, ontleend aan het schematische attributiemodel van Reeder en Brewer (1979), is dat waarnemers positieve competentie-informatie zwaarder wegen dan negatieve, maar negatieve integriteitsinformatie zwaarder wegen dan positieve. Eén fout wordt bij competentie als een uitzondering gezien, terwijl één oneerlijke daad als een betrouwbaar signaal van een gebrekkige integriteit geldt. De schuldbekentenis die in een verontschuldiging vervat zit, is daardoor bij een integriteitsfalen bijzonder schadelijk.

Kim et al. (2006) verfijnden dit beeld in een vervolgstudie waarin niet de keuze tussen verontschuldigen en ontkennen, maar de attributie binnen een verontschuldiging werd gevarieerd (interne versus externe attributie). Zij vonden dat bij een competentiefalen een verontschuldiging met interne attributie (meer schuld op zich nemen, “more blame”) het vertrouwen effectiever herstelde, terwijl bij een integriteitsfalen een verontschuldiging met externe attributie (de schuld deels extern leggen) effectiever bleek, zelfs al zijn waarnemers geneigd situationele verklaringen voor immoreel gedrag te wantrouwen. Een integriteitsschending is met andere woorden zo schadelijk voor vertrouwen dat elke poging tot schuldmitigatie de moeite waard kan zijn. Verwante patronen werden nadien verder onderzocht en genuanceerd (Dirks et al., 2009; Tomlinson et al., 2004). Tegelijk wijzen meta-analyses op contextuele variatie. Het type slachtoffer, de ernst van het incident, de pre-existerende relatie en de geloofwaardigheid van de spreker modereren mee. De openstaande vraag voor deze masterproef is of het interactie-effect ook geldt wanneer (1) de organisatie publiek is, (2) het falen een AI-systeem betreft, en (3) Belgische burgers de respondenten zijn.

2.5.4 Verontschuldigen na AI-falen: opkomende inzichten

De experimentele literatuur over vertrouwensherstel na specifiek AI-falen is recent en nog beperkt. Kim en Song (2021) onderzochten in een investeringssimulatie hoe burgers reageren op verontschuldigen door AI-agenten met verschillende attributies, en vonden dat antropomorfisme van het agent een belangrijke moderator is. Glikson en Woolley (2020) bieden een uitgebreide review van vertrouwen in AI, waaruit blijkt dat de vertrouwensdynamiek tussen mens en AI zowel overeenkomsten als verschillen kent met mens-mens vertrouwen. Aoki (2020) ten slotte toonde aan dat burgers verschillend reageren op AI-chatbots in publieke dienstverlening afhankelijk van het beleidsdomein (bv. opvoedingsondersteuning versus afvalbeheer).

Recentere experimenten bevestigen en verfijnen deze eerste inzichten. Pareek, Velloso en Goncalves (2024) toonden in een reeks mens-AI-beslissingstaken aan dat een verontschuldiging de meest effectieve herstelstrategie was nadat een AI-systeem fouten had gemaakt, terwijl een ontkenning het wantrouwen net versterkte. Voor de publieke sector is vooral het grootschalige Nederlandse survey-experiment van Alon-Barkat et al. (2025) relevant: burgers bleken publieke organisaties niet minder verantwoordelijk te houden voor algoritmische dan voor menselijke discriminatie. Dit wijst erop dat technologische delegatie organisaties in de ogen van burgers niet van hun verantwoordelijkheid ontslaat. Wel schreven burgers de organisatie meer verantwoordelijkheid toe wanneer het algoritme intern was ontwikkeld dan wanneer het door een externe private leverancier was geleverd, een bevinding die rechtstreeks raakt aan de geloofwaardigheid van externe attributies richting leveranciers in verontschuldigen.

Voor de publieke sector specifiek voegt opkomende literatuur enkele nuances toe aan het klassieke model. Ten eerste suggereert Busuioc (2021) dat bij AI-falen burgers een hoger niveau van expliciete erkenning verwachten van de menselijke verantwoordelijken. Het is niet voldoende dat “het algoritme” zich verontschuldigt, de leider moet dat ook doen. Ten tweede tonen experimenten dat externe attributies naar het algoritme zelf door burgers vaak gezien kan worden als verantwoordelijkheidsontvluchting, zeker

wanneer de leider duidelijk de keuze had om al dan niet AI in te zetten (Busuioc, 2021). Ten derde wijst onderzoek over algoritmische transparantie erop dat verontschuldigen die uitleg bevatten, significant effectiever zijn voor vertrouwensherstel (Grimmelikhuijsen, 2023; Schiff et al., 2022). Deze inzichten worden in de vignetmaterialen meegenomen door bij elke conditie een gestandaardiseerde verontschuldiging te formuleren waarin enkel de attributie systematisch varieert.

2.6 Conceptueel kader en hypothesen

Op basis van het voorgaande literatuuroverzicht wordt het volgende conceptueel kader voorgesteld. Een organisationeel AI-falen leidt tot een daling van het burgervertrouwen in de organisatie en haar leider. De omvang van deze daling wordt mee bepaald door het type falen (competentie versus integriteit). Vervolgens hanteert de organisatie een verontschuldiging waarin zij de oorzaak intern of extern attribueert. De effectiviteit van deze attributie voor vertrouwensherstel wordt gemodereerd door het type falen. De verwachting dat een dergelijk interactie-effect bestaat, is ontleend aan Kim et al. (2006). De hier voorspelde richting wijkt evenwel af van hun bevinding in een private sollicitatiecontext (interne attributie effectiever bij competentiefalen, externe bij integriteitsfalen): in een publieke context, waar burgers bij een moreel geladen falen een expliciete erkenning van verantwoordelijkheid verwachten en externe attributies richting het algoritme al snel als verantwoordelijkheidsontwijking worden gepercipieerd (Busuioc, 2021; Coombs, 2007), wordt verwacht dat een interne attributie net bij een integriteitsfalen het meest noodzakelijk is. Het uiteindelijke vertrouwensherstel wordt gemeten als post-manipulatie vertrouwen, gecontroleerd voor relevante covariaten zoals algemeen overheidsvertrouwen en politieke betrokkenheid.

Uit dit kader vloeien drie hypothesen voort:

- H1.** Een AI-gerelateerd integriteitsfalen leidt tot een sterkere daling van burgervertrouwen in de publieke organisatie en haar leider dan een AI-gerelateerd competentiefalen.
- H2.** Een verontschuldiging met interne attributie leidt tot een sterker vertrouwensherstel dan een verontschuldiging met externe attributie.
- H3.** Het effect van de attributie in een verontschuldiging op vertrouwensherstel wordt gemodereerd door het type falen: interne attributie is effectiever bij integriteitsfalen, terwijl externe attributie effectiever (of minder schadelijk) is bij competentiefalen.

Deze hypothesen worden in Hoofdstuk 3 (Methodologie) operationeel ingevuld in een 2 x 2 between-subjects experimenteel design, waarbij het type falen (competentie versus integriteit van het AI-systeem) en het type attributie (intern versus extern) systematisch worden gevarieerd in vier vignetcondities. Het ontwerp laat toe zowel de hoofdeffecten (H1 en H2) als het cruciale interactie-effect (H3) te toetsen.

3. Methodologie

Dit hoofdstuk verantwoordt de methodologische keuzes van het onderzoek. Achtereenvolgens komen de onderzoeksstrategie en het design (3.1), het vignetmateriaal (3.2), de steekproef en dataverzameling (3.3), de operationalisering en meetinstrumenten (3.4) en de analysestrategie (3.5) aan bod.

3.1 Onderzoeksstrategie en -ontwerp

Om de centrale onderzoeksvraag te beantwoorden werd gekozen voor een kwantitatief, experimenteel design in de vorm van een online vignetonderzoek. Het experiment hanteert een 2 (type falen: competentie versus integriteit) x 2 (type attributie: intern versus extern) between-subjects design. Deze keuze sluit aan bij het paradigma van Kim et al. (2006) en laat toe causale uitspraken te doen over het effect van de gemanipuleerde variabelen op het burgersvertrouwen.

Een belangrijke uitbreiding ten opzichte van een klassiek post-only design is dat het vertrouwen op drie momenten werd gemeten: voor het falen (T1, baseline), na de confrontatie met het falen (T2) en na de officiële verontschuldiging (T3). Deze herhaalde meting binnen dezelfde respondent maakt het mogelijk de vertrouwensdaling en het vertrouwensherstel afzonderlijk te modelleren. Bovendien verhoogt een within-subjects component de statistische power, wat bij een beperkte steekproef een aanzienlijk voordeel is.

3.2 Opbouw van het vignetmateriaal

Het vignetmateriaal, opgenomen in bijlage A, is opgebouwd uit drie opeenvolgende tekstonderdelen rond de fictieve Dienst Sociale Zaken van de gemeente Veldburg. Het eerste onderdeel is voor alle respondenten identiek; het tweede en derde onderdeel werden systematisch gevarieerd volgens het 2 × 2 design, wat vier vignetcondities oplevert: competentiefout met interne attributie, competentiefout met externe attributie, integriteitsfout met interne attributie en integriteitsfout met externe attributie.

Het gemeenschappelijke basisscenario introduceert de organisatie en haar leidinggevende. De dienst behandelt jaarlijks ruim vierduizend aanvragen voor sociale bijstand, woonpremies en gemeentelijke tegemoetkomingen. Twee jaar voor het incident voerde directeur dhr. Delvaux een AI-systeem in dat ingediende dossiers analyseert en per aanvraag een aanbeveling formuleert (goedkeuren of weigeren), waarbij medewerkers die aanbeveling kunnen opvolgen of gemotiveerd afwijken. Deze configuratie is bewust gekozen: zij plaatst het systeem in een ondersteunende rol met menselijke eindverantwoordelijkheid, de opzet die in de bestuurspraktijk het meest voorkomt en waarin de attributievraag het scherpst speelt, aangezien het systeem adviseert maar mensen beslissen. Na dit basisscenario volgde de eerste vertrouwensmeting (T1).

Het type falen werd vervolgens gemanipuleerd via een nieuwsbericht. In de competentieconditie onthult een audit dat het systeem door een fout in de rekenlogica en het gebruik van verouderde wettelijke parameters

aanvragen structureel foutief weigerde; het falen wordt gepresenteerd als een technisch kennisgebrek, in neutrale bewoordingen en met de toezegging van terugbetaling. In de integriteitsconditie onthult een intern onderzoek dat het systeem aanvragen uit wijken met een hogere concentratie inwoners met een migratieachtergrond systematisch benadeelde. Hier wordt het falen expliciet moreel geladen, onder meer via het aan slachtofferverenigingen toegeschreven verwijt van “geïnstitutionaliseerde discriminatie” en hun eis tot publieke verantwoording. Om te verzekeren dat uitsluitend de aard van het falen varieerde en niet de ernst of omvang ervan, werden de overige parameters parallel gehouden. In beide versies duurde het falen ruim anderhalf jaar, werden naar schatting 340 burgers ten onrechte een tegemoetkoming geweigerd, kwam het incident aan het licht na analyse door een externe auditor aangesteld door de gemeenteraad, en werd dhr. Delvaux formeel van de bevindingen op de hoogte gebracht. Na dit nieuwsbericht volgde de tweede vertrouwensmeting (T2).

De attributie werd ten slotte gemanipuleerd via een officiële verklaring van dhr. Delvaux. In de interne-attributieconditie neemt hij de verantwoordelijkheid ondubbelzinnig op zich. In de externe-attributieconditie legt hij de oorzaak bij de externe technologieleverancier die het systeem leverde en configureerde, stelt hij dat het systeem niet functioneerde zoals contractueel overeengekomen, en kondigt hij juridische stappen tegen die leverancier aan. Beide verklaringen bevatten dezelfde constante elementen: een rechtstreekse aanspreking van de getroffen burgers, een expliciete verontschuldiging, de erkenning van de geleden schade en de toezegging dat alle onterecht geweigerde tegemoetkomingen met terugwerkende kracht worden rechtgezet. Zij zijn bovendien van vergelijkbare lengte en register, zodat de attributie van de oorzaak het enige systematisch gevarieerde element is. Na deze verklaring volgde de derde vertrouwensmeting (T3).

Op elk van de drie meetmomenten werd het vertrouwen in zowel de organisatie als de leidinggevende gemeten met identieke itemsets, opgesteld in lijn met de dimensies van Mayer et al. (1995) (zie 3.4.1). Het begrip van beide manipulaties werd na afloop gecontroleerd via manipulatiechecks (zie 3.4.2).

3.3 Steekproef en dataverzameling

De data werden verzameld via een online vragenlijst in Qualtrics, verspreid via sociale media en het persoonlijke netwerk. De inclusiecriteria waren Belgische burgers van achttien jaar of ouder. De originele dataset bestond uit 140 beëindigde enquêtes, met een verdeling over de condities van 39, 30, 34 en 37 respondenten. Na toepassing van de manipulatiechecks en verwijdering van een onbruikbare observatie resteerde een analysesteekproef van 93 respondenten. De verdeling over de vier condities bedroeg dan 33, 16, 26 en 18 respondenten.

De steekproefgrootte is beperkt voor het detecteren van kleine tot middelgrote effecten. Dit wordt erkend als een beperking en wordt meegenomen in de interpretatie door bij elke toets de effectgrootte en de geobserveerde power te rapporteren.

3.4 Operationalisering en meetinstrumenten

3.4.1 *Afhankelijke variabele: burgersvertrouwen*

Burgersvertrouwen werd gemeten met een gevalideerde schaal gebaseerd op Grimmelikhuijsen en Knies (2017) en Mayer et al. (1995), met items voor de dimensies vermogen, welwillendheid en integriteit. Het vertrouwen werd afzonderlijk bevraagd voor de organisatie en voor de leidinggevende, telkens met vier items op een zevenpunts Likertschaal. Per meetmoment werd een composietscore berekend als het gemiddelde van de vier items, wat resulteerde in zes afhankelijke variabelen.

3.4.2 *Onafhankelijke variabelen en manipulatiechecks*

De twee onafhankelijke variabelen, type falen en type attributie, werden gemanipuleerd via de vier vignetcondities. Manipulatiechecks gingen na of respondenten het type falen correct identificeerden en of zij de attributie als intern dan wel extern percipieerden. Daarnaast werd een attentiecheck opgenomen over de identiteit van de leidinggevende. Respondenten die minstens een van beide manipulatiechecks faalden, werden uitgesloten uit de analyses, aangezien een correcte perceptie van beide onafhankelijke variabelen een voorwaarde is voor de validiteit van het design.

3.4.3 *Controlevariabelen*

Naast de experimentele variabelen werden geslacht, leeftijd en hoogst behaalde diploma bevraagd, alsook de mate van betrokkenheid bij overheidsbeslissingen en de vertrouwdheid met het gebruik van AI-systemen door overheden. De demografische kenmerken dienen ter beschrijving van de steekproef en ter controle van de randomisatie. De twee laatstgenoemde variabelen werden bovendien als covariaten opgenomen in een aanvullende analyse.

3.5 Analysestrategie

De analyse verliep in vier stappen. Eerst werd de betrouwbaarheid van de zes vertrouwensschalen nagegaan via Cronbach's alfa. Vervolgens werden de steekproefsamenstelling en de randomisatie gecontroleerd. De kern van de analyse bestond uit een mixed repeated-measures ANOVA met tijd als within-subjects factor en type attributie en type falen als between-subjects factoren, afzonderlijk uitgevoerd voor organisatie- en leidersvertrouwen. Ten slotte werd deze analyse herhaald met de covariaten als robuustheidscontrole (Van Thiel, 2021).

Een belangrijk methodologisch aandachtspunt verdient hier expliciete vermelding. Zoals in hoofdstuk 4 wordt aangetoond, verschillen de condities reeds op T1, dus op een moment waarop nog geen manipulatie heeft plaatsgevonden. Een rechtstreekse vergelijking van de absolute vertrouwensniveaus op T2 en T3 tussen condities is daardoor moeilijk interpreteerbaar, omdat de startposities niet gelijk zijn. Om die reden wordt in de rapportering primair gefocust op de veranderingen binnen elke conditie, namelijk de daling van T1 naar T2 en het herstel van T2 naar T3. Deze binnen-conditie veranderingen zijn immuun voor baselineverschillen

en vormen daarmee de meest valide basis voor interpretatie. De interactie-effecten met tijd in de repeated-measures ANOVA toetsen precies deze veranderingspatronen en zijn om die reden de centrale toetsingsstatistieken van dit onderzoek.

Om dezelfde reden werd afgezien van een post-hoc decompositie van het between-subjects interactie-effect tussen attributie en type falen. Dergelijke vervolganalyses zijn bedoeld om een significant en valide interactie-effect verder te ontleden. Aangezien het effect het gemiddelde vertrouwensniveau over de drie meetmomenten heen betreft en daardoor vertekend wordt door de ongelijke startposities, zou een decompositie ervan geen valide bijkomende inzichten opleveren. De toetsing van de hypothesen steunt bijgevolg uitsluitend op de interactie-effecten met tijd.

4. Resultaten

Dit hoofdstuk presenteert de onderzoeksresultaten. Conform de richtlijnen voor kwantitatief rapporteren (Van Thiel, 2021) wordt de SPSS-output niet rechtstreeks overgenomen, maar weergegeven in zelf opgestelde tabellen. Achtereenvolgens komen de betrouwbaarheid van de meetinstrumenten (4.1), de steekproefbeschrijving en randomisatiecontrole (4.2), de beschrijvende statistieken en de binnen-conditie veranderingen (4.3), de toetsing van de hypothesen voor organisatievertrouwen (4.4) en leidersvertrouwen (4.5), de controle van de statistische aannames (4.6), de robuustheidsanalyse met covariaten (4.7) en een samenvattend overzicht (4.8) aan bod. Alle analyses zijn gebaseerd op $N = 93$.

4.1 Betrouwbaarheid van de meetinstrumenten

Voor elk van de zes vertrouwensschalen werd de interne consistentie berekend. Tabel 1 toont dat alle schalen ruim boven de gehanteerde drempelwaarde van .70 uitkomen, met waarden tussen .887 en .947. Dit wijst op een uitstekende interne consistentie en rechtvaardigt het gebruik van gemiddelde composietscores.

Tabel 1. Interne consistentie van de vertrouwensschalen (Cronbach's alfa)

Schaal	Meetmoment	Aantal items	Cronbach's α	N
Vertrouwen in de organisatie	T1 (baseline)	4	,895	92
Vertrouwen in de organisatie	T2 (na fout)	4	,887	93
Vertrouwen in de organisatie	T3 (na verontschuldiging)	4	,942	93
Vertrouwen in de leidinggevende	T1 (baseline)	4	,947	92
Vertrouwen in de leidinggevende	T2 (na fout)	4	,940	93
Vertrouwen in de leidinggevende	T3 (na verontschuldiging)	4	,940	93

4.2 Steekproefbeschrijving en randomisatiecontrole

De analysesteekproef bestaat uit 93 respondenten, waarvan 49 mannen (52,7%) en 44 vrouwen (47,3%). Wat het opleidingsniveau betreft heeft 24,7% een diploma secundair onderwijs, 30,1% een professionele bachelor, 14,0% een academische bachelor en 31,2% een master of hoger. De leeftijd van de respondenten varieert van 18 tot 82 jaar, met een duidelijke concentratie in twee leeftijdsgroepen: jongvolwassenen tussen 22 en 23 jaar en een groep tussen 42 en 58 jaar.

Om na te gaan of de random toewijzing tot vergelijkbare groepen heeft geleid, werd voor elk achtergrondkenmerk getest of het verschilt tussen de vier condities. Tabel 2 toont dat geen van deze testen significant is, wat erop wijst dat de condities gebalanceerd zijn wat betreft geslacht, opleidingsniveau en leeftijd.

Tabel 2. Randomisatiecontrole: verdeling van achtergrondkenmerken over de vier condities

Kenmerk	Toets	Toetsstatistiek	Df	p	Conclusie
Geslacht	Chi-kwadraat	$\chi^2 = 6,06$	3	,109	Gebalanceerd
Diploma	Chi-kwadraat	$\chi^2 = 6,83$	9	,655	Gebalanceerd
Leeftijd	One-way ANOVA	F = 1,54	3; 89	,209	Gebalanceerd

Bij de chi-kwadraattoets voor diploma dient opgemerkt dat 50% van de cellen een verwachte frequentie onder vijf heeft, wat inherent is aan de combinatie van vier opleidingscategorieën met vier condities bij een steekproef van deze omvang. De uitkomst is daardoor indicatief eerder dan strikt toetsend, maar geeft samen met de andere twee toetsen geen aanwijzing voor een scheve verdeling.

4.3 Beschrijvende statistieken en binnen-conditie veranderingen

Tabel 3 geeft de gemiddelden en standaardafwijkingen van het organisatie- en leidersvertrouwen per conditie en per meetmoment weer, gemeten op een zevenpunts Likertschaal.

Tabel 3. Gemiddelden en standaardafwijkingen van burgervertrouwen per conditie en meetmoment (N = 93)

Conditie	n	T1 baseline	T2 na fout	T3 na verontsch.
Vertrouwen in de organisatie				
Extern + Competentie	33	4,02 (1,25)	2,89 (1,23)	3,61 (1,22)
Extern + Integriteit	16	4,97 (1,22)	3,72 (1,50)	4,61 (1,40)
Intern + Competentie	26	4,25 (1,59)	3,14 (1,51)	4,26 (1,28)
Intern + Integriteit	18	4,26 (1,40)	2,60 (1,08)	4,25 (1,44)
Vertrouwen in de leidinggevende				
Extern + Competentie	33	3,92 (1,28)	2,86 (1,39)	3,58 (1,45)
Extern + Integriteit	16	4,97 (1,44)	3,75 (1,68)	4,42 (1,75)
Intern + Competentie	26	4,13 (1,63)	3,08 (1,40)	4,57 (1,53)
Intern + Integriteit	18	4,00 (1,34)	2,83 (1,20)	4,74 (1,24)

Uit Tabel 3 blijkt onmiddellijk het methodologische aandachtspunt dat in sectie 3.5 werd aangekondigd. De baselinewaarden op T1 lopen uiteen tussen de condities, met een verschil van bijna een volledig schaalpunt tussen de laagste en de hoogste conditie. Omdat op T1 nog geen enkele manipulatie had plaatsgevonden,

kunnen deze verschillen niet aan het experiment worden toegeschreven maar zijn zij het gevolg van toevallige variatie in de toewijzing. Een rechtstreekse vergelijking van de absolute niveaus op T2 en T3 tussen condities zou daardoor misleidend zijn.

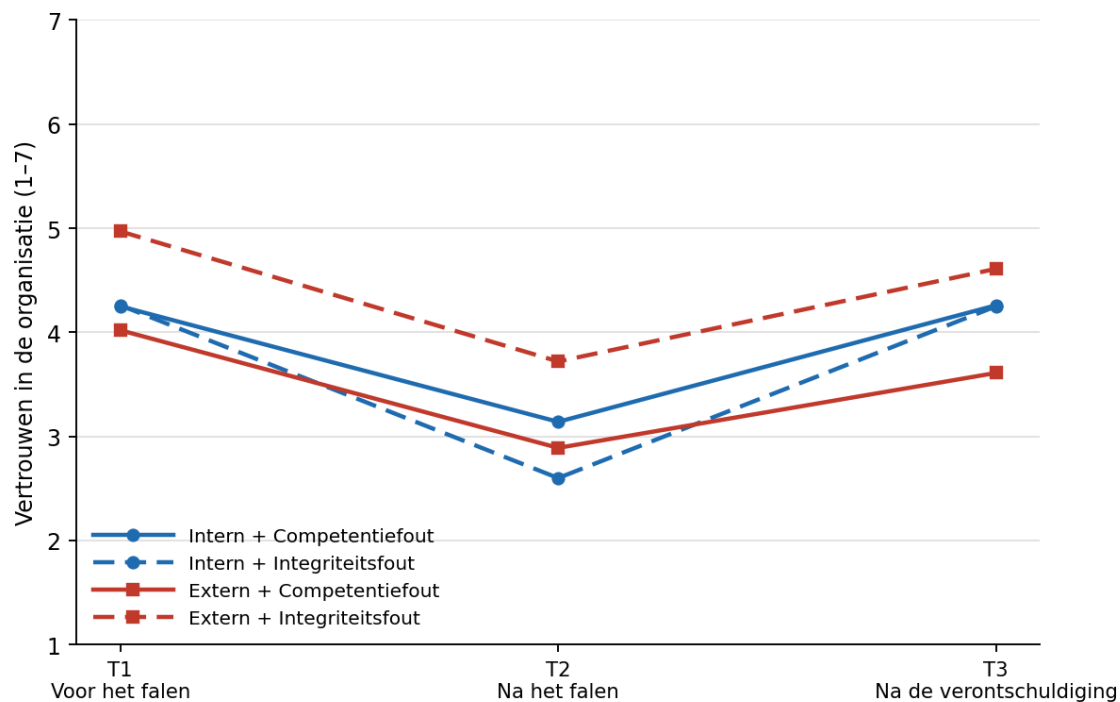
Om die reden worden in Tabel 4 de veranderingen binnen elke conditie weergegeven. Deze veranderingsmaten zijn per definitie ongevoelig voor verschillen in startpositie en vormen daarmee de meest valide basis voor interpretatie. Naast de absolute daling en het absolute herstel wordt ook de herstelratio gerapporteerd, die aangeeft welk aandeel van het verloren vertrouwen door de verontschuldiging wordt teruggewonnen.

Tabel 4. Binnen-conditie veranderingen in burgervertrouwen (N = 93)

Conditie	Daling T1→T2	Herstel T2→T3	Netto T1→T3	Herstelratio
Vertrouwen in de organisatie				
Extern + Competentie	-1,14	+0,72	-0,42	63%
Extern + Integriteit	-1,25	+0,89	-0,36	71%
Intern + Competentie	-1,11	+1,12	+0,01	101%
Intern + Integriteit	-1,67	+1,65	-0,01	99%
Vertrouwen in de leidinggevende				
Extern + Competentie	-1,07	+0,73	-0,34	68%
Extern + Integriteit	-1,22	+0,67	-0,55	55%
Intern + Competentie	-1,06	+1,49	+0,43	141%
Intern + Integriteit	-1,17	+1,90	+0,74	163%

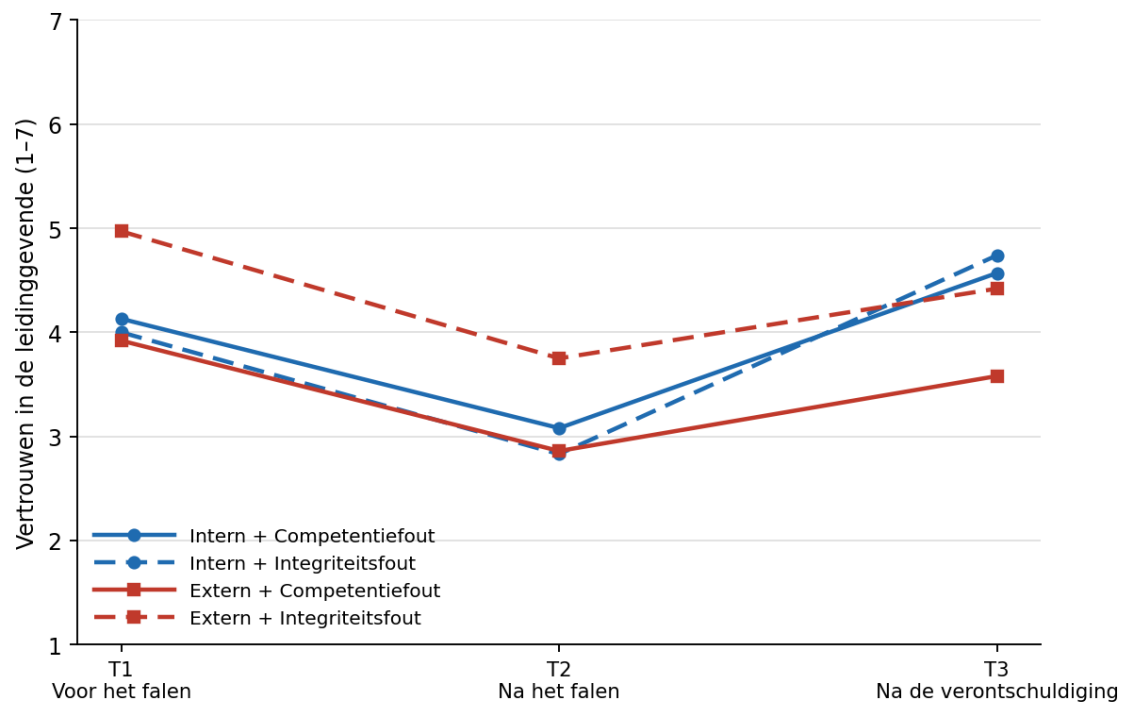
Tabel 4 toont een opvallend consistent patroon. De daling van het vertrouwen na het falen is in alle condities vergelijkbaar en schommelt rond een schaalpunt, ongeacht het type falen of de attributie. Dit is logisch, aangezien de attributie op dat moment nog niet was gecommuniceerd. Het herstel daarentegen verschilt sterk. In de condities met externe attributie wordt slechts 55 tot 71 procent van het verloren vertrouwen teruggewonnen, waardoor het vertrouwen op T3 blijvend onder het beginniveau blijft. In de condities met interne attributie wordt het verlies daarentegen volledig gecompenseerd: bij de organisatie bedraagt het herstel 99 tot 101 procent, bij de leidinggevende zelfs 141 tot 163 procent, wat betekent dat het vertrouwen in de leidinggevende na de verontschuldiging hoger ligt dan voor het incident.

Figuur 1 en Figuur 2 visualiseren dit verloop voor respectievelijk het organisatie- en het leidersvertrouwen. De kleur van de lijnen onderscheidt de attributie (blauw = intern, rood = extern), de lijnstijl het type falen (doorlopend = competentiefout, gestippeld = integriteitsfout).



Figuur 1. Verloop van het vertrouwen in de organisatie per conditie over de drie meetmomenten (N = 93)

Figuur 1 maakt drie zaken zichtbaar. Ten eerste vertonen alle vier de condities het V-vormige verloop van daling en herstel. Ten tweede zijn de ongelijke startposities op T1 duidelijk herkenbaar, in het bijzonder de hogere baseline van de conditie extern-integriteitsfout. Ten derde valt op dat de twee blauwe lijnen (interne attributie) op T3 terugkeren naar hun eigen beginniveau, terwijl de twee rode lijnen (externe attributie) daar zichtbaar onder blijven steken.



Figuur 2. Verloop van het vertrouwen in de leidinggevende per conditie over de drie meetmomenten (N = 93)

In Figuur 2 is het interactie-effect tussen tijd en attributie visueel het meest sprekend. Waar de vier condities op T2 nog dicht bij elkaar liggen, waaieren de lijnen op T3 sterk uit volgens de attributie: de blauwe lijnen stijgen tot boven hun eigen baseline en eindigen als hoogste van alle condities, terwijl de rode lijnen onder hun startpunt blijven. De lijnen kruisen elkaar tussen T2 en T3, wat de kern van het significante Tijd × Attributie-effect grafisch weergeeft: niet de startpositie maar de attributie in de verontschuldiging bepaalt waar het vertrouwen eindigt.

Dit patroon vormt de kern van de bevindingen en wordt in de volgende secties statistisch getoetst via de interactie-effecten met tijd.

4.4 Toetsing van de hypothesen: vertrouwen in de organisatie

Tabel 5 toont de resultaten van de mixed repeated-measures ANOVA voor het organisatievertrouwen. Aangezien Mauchly's toets niet significant is (zie sectie 4.6), worden de waarden onder de aanname van sfericiteit gerapporteerd.

Tabel 5. Repeated-measures ANOVA voor vertrouwen in de organisatie: within-subjects effecten (1) en between-subjects effecten (2)

Within-subjects Effecten	F	df	p	η^2	Power
Tijd	67,44	2; 178	< ,001	,431	1,00
Tijd × Attributie	3,03	2; 178	,051	,033	,58
Tijd × Type falen	1,39	2; 178	,251	,015	,30
Tijd × Attributie × Type falen	0,50	2; 178	,610	,006	,13

Between-subjects Effecten	F	df	p	η^2	Power
Hoofdeffect Attributie	,47	1; 89	,495	,005	,10
Hoofdeffect Falen	2,14	1; 89	,147	,024	,30
Attributie × Type falen	4,72	1; 89	,032	,050	,57

Het hoofdeffect van tijd is bijzonder sterk ($F(2,178) = 67,44$, $p < ,001$, $\eta^2 = ,431$, power = 1,00). Dit bevestigt dat het vertrouwen significant verandert over de drie meetmomenten. De contrastanalyse verduidelijkt de vorm van die verandering: de lineaire component is niet significant ($F(1,89) = 2,40$, $p = ,125$) terwijl de kwadratische component zeer sterk significant is ($F(1,89) = 147,70$, $p < ,001$, $\eta^2 = ,624$). Statistisch bevestigt dit het V-vormige verloop van daling gevolgd door herstel, en niet een eenvoudige neerwaartse of opwaartse trend.

Het interactie-effect tussen tijd en attributie is randsignificant ($F(2,178) = 3,03$, $p = ,051$, $\eta^2 = ,033$). De kwadratische component van deze interactie benadert eveneens de significantiedrempel ($F(1,89) = 3,87$, $p = ,052$, $\eta^2 = ,042$). Dit wijst erop dat het herstelpatroon voor de organisatie verschilt naargelang de attributie, in de richting die Tabel 4 laat zien, maar dat dit verschil bij de huidige steekproefomvang net niet de conventionele drempel haalt. Gezien de geobserveerde power van slechts ,58 voor dit effect is een type-II-fout hier een reële mogelijkheid.

De interacties met het type falen zijn niet significant, noch afzonderlijk ($p = ,251$) noch in combinatie met de attributie ($p = ,610$). Het veranderingspatroon van het organisatievertrouwen verschilt dus niet naargelang het falen als een competentie- dan wel als een integriteitskwestie werd gepresenteerd.

Tussen de respondenten onderling is er wel een significant interactie-effect tussen attributie en type falen ($F(1,89) = 4,72$, $p = ,032$, $\eta^2 = ,050$). Gezien de vastgestelde baselineverschillen moet dit between-subjects effect echter met de nodige voorzichtigheid geïnterpreteerd worden: het betreft het gemiddelde

vertrouwensniveau over de drie meetmomenten heen, waarin de ongelijke startposities doorwerken. Het effect is bijgevolg indicatief eerder dan conclusief.

4.5 Toetsing van de hypothesen: vertrouwen in de leidinggevende

Tabel 6 toont de overeenkomstige resultaten voor het leidersvertrouwen. Hier tekent zich een duidelijker beeld af.

Tabel 6. Repeated-measures ANOVA voor vertrouwen in de leidinggevende: within-subjects effecten (1) en between-subject effecten (2)

Within-subjects effecten	F	df	p	η^2	Power
Tijd	52,25	2; 178	< ,001	,370	1,00
Tijd × Attributie	9,88	2; 178	< ,001	,100	,98
Tijd × Type falen	0,25	2; 178	,782	,003	,09
Tijd × Attributie × Type falen	0,58	2; 178	,562	,006	,14

Between-subjects effecten	F	df	p	η^2	Power
Hoofdeffect Attributie	,09	1; 89	,924	,000	,05
Hoofdeffect Falen	2,47	1; 89	,120	,027	,34
Attributie × Type falen	3,34	1; 89	,071	,036	,44

Ook hier is het hoofdeffect van tijd sterk significant ($F(2,178) = 52,25$, $p < ,001$, $\eta^2 = ,370$), met opnieuw een dominante kwadratische component ($F(1,89) = 111,74$, $p < ,001$, $\eta^2 = ,557$) tegenover een niet-significante lineaire component ($p = ,606$).

Het centrale resultaat van dit onderzoek is het interactie-effect tussen tijd en attributie, dat voor de leidinggevende wel duidelijk significant is ($F(2,178) = 9,88$, $p < ,001$, $\eta^2 = ,100$, power = ,98). Het gaat om een middelgroot effect dat met hoge zekerheid is vastgesteld. Zowel de lineaire ($F(1,89) = 14,31$, $p < ,001$, $\eta^2 = ,139$) als de kwadratische component ($F(1,89) = 4,82$, $p = ,031$, $\eta^2 = ,051$) van deze interactie is significant, wat aangeeft dat de attributie zowel de eindpositie als de vorm van het herstelverloop beïnvloedt.

De inhoudelijke betekenis van dit effect blijkt uit Tabel 4. Wanneer de leidinggevende de verantwoordelijkheid intern attribueert, herstelt het vertrouwen niet alleen volledig maar overstijgt het zelfs het oorspronkelijke niveau, met een herstelratio van 141 tot 163 procent. Bij een externe attributie blijft het herstel beperkt tot 55 tot 68 procent en blijft het vertrouwen structureel beschadigd. Het persoonlijk opnemen van verantwoordelijkheid door een leidinggevende blijkt dus niet alleen schadebeperkend maar potentieel vertrouwenversterkend.

De interacties met het type falen zijn opnieuw niet significant ($p = ,782$ respectievelijk $p = ,562$). Het between-subjects interactie-effect tussen attributie en type falen is randsignificant ($F(1,89) = 3,34$, $p = ,071$, $\eta^2 = ,036$), maar valt onder hetzelfde interpretatievoorbehoud als hierboven beschreven.

4.6 Controle van de statistische aannames

De aannames voor de repeated-measures ANOVA werden systematisch gecontroleerd. Tabel 7 vat de resultaten samen. Geen enkele aanname wordt geschonden, wat betekent dat de gerapporteerde toetsen zonder correctie geïnterpreteerd mogen worden.

Tabel 7. Controle van de statistische aannames

Aanname	Toets	Organisatie	Leidinggevende	Conclusie
Sfericiteit	Mauchly's W	$W = ,983$; $p = ,460$	$W = ,986$; $p = ,542$	Voldaan
Gelijkheid covariantiematrices	Box's M	$M = 19,44$; $p = ,448$	$M = 8,44$; $p = ,981$	Voldaan
Homogeniteit varianties T1	Levene	$F = 1,23$; $p = ,302$	$F = 0,98$; $p = ,406$	Voldaan
Homogeniteit varianties T2	Levene	$F = 1,11$; $p = ,349$	$F = 1,43$; $p = ,240$	Voldaan
Homogeniteit varianties T3	Levene	$F = 0,28$; $p = ,842$	$F = 0,77$; $p = ,512$	Voldaan

4.7 Robuustheidsanalyse met covariaten

Als robuustheidscontrole werd de volledige analyse herhaald met de betrokkenheid bij overheidsbeslissingen en de vertrouwdsheid met AI-gebruik door overheden als covariaten. Tabel 8 vergelijkt de centrale effecten uit beide modellen.

Tabel 8. Vergelijking van de hoofdanalyse met het model inclusief covariaten

Effect	Zonder covariaten	Met covariaten	Conclusie
Tijd (organisatie)	$F = 67,44$; $p < ,001$	$F = 18,59$; $p < ,001$	Stabiel
Tijd $\times$ Attributie (organisatie)	$F = 3,03$; $p = ,051$	$F = 2,42$; $p = ,092$	Vergelijkbaar
Tijd (leidinggevende)	$F = 52,25$; $p < ,001$	$F = 15,26$; $p < ,001$	Stabiel
Tijd $\times$ Attributie (leidinggevende)	$F = 9,88$; $p < ,001$	$F = 9,00$; $p < ,001$	Stabiel

Geen van beide covariaten oefent een significant effect uit op het vertrouwensverloop. Voor betrokkenheid geldt $p = ,852$ bij de organisatie en $p = ,752$ bij de leidinggevende; voor AI-vertrouwdsheid respectievelijk $p = ,078$ en $p = ,155$. De centrale bevinding, namelijk het significante interactie-effect tussen tijd en attributie voor het leidersvertrouwen, blijft onverkort overeind na controle voor deze achtergrondkenmerken. Dit versterkt het vertrouwen in de robuustheid van de resultaten.

4.8 Samenvattend overzicht van de hypothesetoetsing

Tabel 9. Overzicht van de hypothesetoetsing

Hypothese	Organisatie	Leidinggevende
H1 – Een integriteitsfout leidt tot een grotere daling van het burgersvertrouwen dan een competentiefout	Niet bevestigd $p = ,251$	Niet bevestigd $p = ,782$
H2 – Een verontschuldiging met interne attributie leidt tot meer vertrouwensherstel dan externe attributie	Randsignificant $p = ,051$	Bevestigd $p < ,001$; $\eta^2 = ,100$
H3 – Het effect van de attributie wordt gemodereerd door het type falen	Niet bevestigd $p = ,610$	Niet bevestigd $p = ,562$

Samenvattend leveren de analyses drie bevindingen op. Ten eerste is er robuust bewijs dat het burgersvertrouwen significant daalt na een AI-gerelateerd falen en vervolgens significant herstelt na een verontschuldiging. Dit geldt zowel voor de organisatie als voor de leidinggevende en betreft in beide gevallen een groot effect. De contrastanalyses tonen dat het verloop expliciet kwadratisch is, wat het V-vormige patroon van breuk en herstel statistisch onderbouwt.

Ten tweede blijkt de attributie in de verontschuldiging bepalend voor de mate van herstel, maar enkel op een statistisch significante wijze voor het vertrouwen in de individuele leidinggevende. Een interne attributie leidt daar tot volledig en zelfs meer dan volledig herstel, terwijl een externe attributie het vertrouwen blijvend beschadigd laat. Voor de organisatie wijst het patroon in dezelfde richting maar bereikt het effect net niet de conventionele significantiedrempel, wat gezien de beperkte power niet als weerlegging maar als onbeslist moet worden geïnterpreteerd. H2 wordt daarmee voor de leidinggevende bevestigd en voor de organisatie niet.

Ten derde vindt het onderzoek geen ondersteuning voor de verwachting dat het type falen de vertrouwensdynamiek beïnvloedt. Noch de daling na het incident, noch het herstel na de verontschuldiging verschilt significant tussen een competentie- en een integriteitsfout, en er is geen aanwijzing voor het door Kim et al. (2006) gesuggereerde moderatie-effect. H1 en H3 worden dus niet bevestigd.

Deze bevindingen, en in het bijzonder het contrast tussen het duidelijke attributie-effect op leidersniveau en de afwezigheid daarvan op organisatieniveau, vormen het uitgangspunt voor de discussie in het volgende hoofdstuk.

5. Discussie en conclusie

Dit slothoofdstuk interpreteert de bevindingen in het licht van de hypothesen en de bestaande literatuur (5.1 en 5.2), beschrijft de theoretische bijdrage (5.3) en de praktische implicaties (5.4), bespreekt de beperkingen van het onderzoek (5.5), formuleert aanbevelingen voor toekomstig onderzoek (5.6) en sluit af met een antwoord op de centrale onderzoeksvraag (5.7).

5.1 Interpretatie van de belangrijkste bevindingen

De eerste en meest robuuste bevinding is dat burgersvertrouwen zich gedraagt als een dynamisch construct dat breekbaar maar ook herstelbaar is. In alle condities daalde het vertrouwen sterk na het AI-gerelateerde falen en veerde het op na de verontschuldiging, een patroon dat statistisch werd bevestigd door de dominante kwadratische component van het tijdseffect. Dat een enkele officiële communicatie een substantieel deel van het verloren vertrouwen kan terugwinnen, onderstreept het belang van crisiscommunicatie voor publieke organisaties. De vertrouwensbreuk na een incident is geen voldongen feit, maar een situatie waarin de reactie van de organisatie het verschil maakt.

De tweede bevinding betreft de rol van de attributie, en die rol blijkt fundamenteel te verschillen naargelang het object van vertrouwen. Voor het vertrouwen in de individuele leidinggevende was het effect van de attributie groot, significant en robuust: een verontschuldiging waarin de leidinggevende de verantwoordelijkheid intern opnam, leidde tot een herstel van 141 tot 163 procent van het verloren vertrouwen, waardoor het vertrouwen na het incident hoger lag dan ervoor. Bij een externe attributie bleef het herstel steken op 55 tot 68 procent en bleef het vertrouwen structureel beschadigd. Voor het vertrouwen in de organisatie wees het patroon in dezelfde richting, maar bereikte het effect net niet de conventionele significantiedrempel. Gezien de beperkte statistische power moet dit laatste niet als een weerlegging maar als een onbesliste uitkomst worden gelezen.

Een mogelijke verklaring voor dit differentiële patroon ligt in de aard van wat een attributie communiceert. Een verontschuldiging wordt uitgesproken door een persoon, en de wijze waarop die persoon de oorzaak toewijst zegt iets over diens karakter. Wie de fout intern attribueert, toont eerlijkheid en verantwoordelijkheidszin, kwaliteiten die rechtstreeks aansluiten bij de integriteits- en welwillendheidsdimensies van het vertrouwensmodel van Mayer et al. (1995). Die karaktersignalen hechten zich in de eerste plaats aan de spreker zelf. Het vertrouwen in de organisatie als geheel steunt daarentegen sterker op structurele factoren zoals prestaties en transparantie (Grimmelikhuijsen et al., 2013), die door een enkele communicatieve handeling minder direct worden geraakt. De verontschuldiging herstelt met andere woorden vooral de reputatie van wie ze uitspreekt.

De derde bevinding is de afwezigheid van elk effect van het type falen. Noch de daling na het incident, noch het herstel na de verontschuldiging verschilden significant tussen de competentie- en de integriteitsconditie,

en het verwachte moderatie-effect op de werking van de attributie bleef eveneens uit. De daling van het vertrouwen bleek in alle condities opvallend uniform, ongeveer een schaalpunt, wat suggereert dat burgers in de eerste plaats reageerden op het feit dat een kwetsbare burger door een overheidsdienst werd benadeeld, en minder op de precieze aard van de onderliggende fout.

5.2 Vergelijking met de bestaande literatuur

De bevindingen nuanceren het invloedrijke werk van Kim et al. (2004) op twee punten. Ten eerste vond dit onderzoek geen bevestiging voor de stelling dat integriteitsfalen een sterkere en hardnekkiger vertrouwensbreuk veroorzaakt dan competentiefalen (Kim et al., 2004; Dirks et al., 2009). Verschillende verklaringen zijn denkbaar. De AI-context kan het onderscheid tussen beide faaltypes vertroebeld hebben: wanneer een algoritme een beslissing neemt die foutloopt, kunnen burgers het incident als een technisch gebrek percipiëren, ongeacht of het scenario de oorzaak als onkunde dan wel als onethisch handelen presenteerde. Daarnaast is het mogelijk dat de verwachtingen van burgers tegenover publieke organisaties anders gestructureerd zijn dan die van de studenten en werknemers in het oorspronkelijke onderzoek, dat zich situeerde in een sollicitatiecontext in de private sector. De publieke sector wordt door burgers mogelijk als een geheel beoordeeld waarin het onderscheid tussen kunnen en willen minder opvallend is.

Ten tweede werd het door Kim et al. (2004, 2006) gedocumenteerde matching-effect, waarbij de optimale herstelstrategie afhangt van het type falen, niet gerepliceerd. Het design van dit onderzoek sluit daarbij het dichtst aan bij Kim et al. (2006), die eveneens twee varianten van een verontschuldiging vergeleken (interne versus externe attributie) en in een private sollicitatiecontext vonden dat interne attributie effectiever was bij competentiefalen en externe attributie bij integriteitsfalen. In de huidige publieke AI-context bleek daarentegen geen enkel moderatie-effect en was interne attributie over de faaltypes heen de effectievere strategie (althans voor het vertrouwen in de leidinggevende). Dit suggereert dat de attributieschema's die het matching-effect in interpersoonlijke en private settings aandrijven, niet zonder meer overdraagbaar zijn naar de relatie tussen burger en overheid. De hogere morele verwachtingen tegenover publieke organisaties kunnen het opnemen van verantwoordelijkheid tot de dominante strategie maken, ongeacht het type falen. De niet-replicatie weerlegt het oorspronkelijke effect dus niet, maar bakent de reikwijdte ervan af.

De bevinding dat interne attributie het herstel bevordert, sluit daarentegen naadloos aan bij de attributietheorie van Weiner (1986). Een interne, controleerbare attributie door de actor zelf ontwapent de toeschrijving van schuld door het publiek en maakt ruimte voor sympathie en hernieuwd vertrouwen, terwijl een externe attributie als ontwijkend kan worden ervaren en de boosheid in stand houdt. Dat dit mechanisme in de huidige studie het sterkst doorwerkte op het vertrouwen in de leidinggevende, verfijnt de toepassing van de theorie op organisationele contexten: de attributie werkt via de perceptie van de attribuerende actor.

5.3 Theoretische bijdrage

Dit onderzoek levert een bijdrage aan vier onderzoekslijnen die tot op heden grotendeels los van elkaar zijn ontwikkeld: de literatuur over vertrouwensherstel, het onderzoek naar institutioneel vertrouwen in de publieke sector, de attributietheorie en de opkomende literatuur over algoritmische besluitvorming door overheden.

De eerste bijdrage is de overdracht van het experimentele herstelparadigma naar een publieke sectorcontext met burgers als vertrouwensgevers. Het onderzoek van Kim et al. (2004, 2006) en de conceptuele uitwerking van Dirks et al. (2009) situeren zich hoofdzakelijk in interpersoonlijke en private organisationele relaties, waarin de vertrouwensgever doorgaans over exit-opties beschikt en waarin de relatie vrijwillig is. Burgers bevinden zich in een fundamenteel andere positie: zij kunnen de overheidsdienst die hen benadeelt niet verlaten en blijven ervan afhankelijk. Dat dit verschil ertoe doet, blijkt uit de bevinding dat het door Kim et al. gedocumenteerde matching-effect zich hier niet reproduceert. Het onderzoek bakent daarmee de reikwijdte van een invloedrijk theoretisch mechanisme af en toont aan dat de assumptie van overdraagbaarheid tussen sectoren niet zonder meer houdbaar is.

De tweede bijdrage betreft het onderscheid tussen het object van vertrouwen. Door organisatie- en leidersvertrouwen gelijktijdig maar afzonderlijk te meten, wordt zichtbaar dat een identieke herstelboodschap op beide niveaus verschillend uitwerkt: de attributie beïnvloedde het herstel van het leidersvertrouwen substantieel, terwijl het organisatievertrouwen zich daar grotendeels ongevoelig voor toonde. In veel bestaand onderzoek worden beide vormen van vertrouwen samengevoegd of wordt er slechts een gemeten, waardoor dit onderscheid onzichtbaar blijft. De bevinding sluit aan bij en verfijnt het model van Mayer et al. (1995): de dimensies welwillendheid en integriteit, die door een verontschuldiging het meest rechtstreeks worden aangesproken, blijken sterker gebonden aan de persoon die de boodschap uitspreekt dan aan de organisatie die zij vertegenwoordigt. Voor het organisatievertrouwen lijken structurele antecedenten zoals prestatie en transparantie (Grimmelikhuijsen et al., 2013) doorslaggevender dan communicatieve handelingen.

De derde bijdrage situeert zich binnen de attributietheorie. Weiner (1986) beschrijft attributie primair als een cognitief proces bij de waarnemer, die een oorzaak toewijst en op basis daarvan emotioneel reageert. Dit onderzoek benadert attributie daarentegen als een communicatieve handeling van de actor zelf, en toont dat de effecten daarvan asymmetrisch neerslaan: zij hechten zich in de eerste plaats aan de spreker. Wie publiek verantwoordelijkheid opneemt, zendt daarmee een signaal over het eigen karakter dat losstaat van de organisatie waarvoor hij spreekt. Deze verschuiving van de waarnemer naar de attribuerende actor als analytisch aangrijpingspunt vormt een uitbreiding van het theoretisch kader die in vervolgonderzoek verder kan worden getoetst.

De vierde bijdrage is contextueel en methodologisch. Inhoudelijk verbindt het onderzoek de herstelliteratuur met de vraag naar verantwoordelijkheid bij algoritmische besluitvorming, een thema dat aan urgentie wint

naarmate overheden AI-systemen inzetten in besluitvormingsprocessen (Busuioc, 2021; Alon-Barkat et al., 2025). Waar die literatuur zich tot nog toe vooral richt op de vraag wie burgers verantwoordelijk achten, toont dit onderzoek wat er gebeurt wanneer de organisatie die vraag zelf beantwoordt. Methodologisch onderscheidt de studie zich van het gangbare post-only design door het vertrouwen op drie momenten te meten binnen dezelfde respondent. Die opzet maakt het mogelijk de vertrouwensbreuk en het vertrouwensherstel analytisch te scheiden, en maakt bovendien een maat als de herstelratio beschikbaar, die uitdrukt welk aandeel van het verloren vertrouwen wordt teruggewonnen. Juist doordat deze maat het herstel afzet tegen de eigen startpositie van elke conditie, is zij ongevoelig voor baselineverschillen. Dit onderzoek illustreert daarmee ook de meerwaarde van herhaalde metingen in vignetexperimenten over vertrouwen.

5.4 Praktische implicaties

Voor de praktijk van publieke crisiscommunicatie zijn de implicaties opvallend concreet. De resultaten suggereren dat een leidinggevende die na een incident persoonlijk en ondubbelzinnig de verantwoordelijkheid opneemt, het eigen vertrouwenskapitaal niet alleen kan beschermen maar zelfs kan versterken. De reflex om de oorzaak buiten de eigen organisatie te leggen, bijvoorbeeld bij een externe leverancier of bij het AI-systeem zelf, blijkt daarentegen contraproductief: het vertrouwen blijft dan structureel onder het oorspronkelijke niveau. Dit laatste is in een AI-context bijzonder relevant. Het afschuiven van de schuld op het algoritme kan door burgers als ontwijkend worden ervaren, precies omdat de organisatie zelf heeft gekozen om dat systeem in te zetten. Voor overheidsorganisaties die algoritmische besluitvorming implementeren, is de les dat de verantwoordelijkheid voor de uitkomsten daarvan communicatief niet delegeerbaar is.

Deze les sluit rechtstreeks aan bij het juridische kader dat in sectie 2.2.1 werd geschetst. De Europese AI Act classificeert AI-toepassingen in de sociale zekerheid als hoog-risico en verplicht tot menselijke supervisie en aantoonbare verantwoordingsstructuren (Europese Commissie, 2024). De bevindingen van dit onderzoek geven die juridische verplichting een communicatieve pendant: menselijke supervisie is niet alleen een organisatorische vereiste vooraf, maar moet na een incident ook zichtbaar worden gemaakt in wie de verantwoordelijkheid publiek opneemt. Een organisatie die formeel aan de supervisievereisten voldoet maar na een falen naar het systeem of de leverancier wijst, ondergraaft communicatief wat zij juridisch geacht wordt te garanderen.

De implicaties reiken bovendien verder dan de individuele leidinggevende. Voor toezichthouders en gemeenteraden suggereren de resultaten dat de attributiekeuze in crisiscommunicatie een relevant criterium is bij de beoordeling van incidentafhandeling. Voor leveranciers van AI-systemen impliceren zij dat contractuele aansprakelijkheid de publieke verantwoordelijkheid van de afnemende organisatie niet vervangt. Zelfs wanneer de fout aantoonbaar bij de leverancier ligt, schaadt het publiek afschuiven ervan het vertrouwen, wat pleit voor een scheiding tussen juridisch verhaal achter de schermen en publieke

verantwoording voor de schermen. Voor burgers en hun vertegenwoordigers ten slotte biedt het onderzoek een concreet criterium om overheidscommunicatie na algoritmische incidenten kritisch te beoordelen.

Bij dit alles hoort een normatieve kanttekening. De strategisch effectiefste aanbeveling, als leidinggevende steeds persoonlijk de verantwoordelijkheid opnemen, ook wanneer de feitelijke oorzaak deels of geheel extern ligt, staat op gespannen voet met het beginsel dat overheidscommunicatie waarheidsgetrouw hoort te zijn. Een interne attributie die de feitelijke oorzakelijkheid geweld aandoet, kan het vertrouwen op korte termijn versterken, maar berust dan zelf op een vorm van misleiding en is bij latere onthulling mogelijk schadelijker dan een eerlijke externe attributie. De aanbeveling moet daarom worden gelezen als een pleidooi voor het opnemen van systeemverantwoordelijkheid: de erkenning dat de organisatie zelf koos voor de inzet van het systeem en instond voor het toezicht erop. Die erkenning is ook bij een technisch externe oorzaak feitelijk correct, en zij is iets anders dan een vrijgeleide om oorzaken te verdraaien.

5.5 Beperkingen van het onderzoek

De bevindingen moeten gelezen worden binnen de grenzen van het onderzoeksdesign. Een eerste beperking betreft de steekproef. Met 93 respondenten na uitsluiting is de statistische power voor het detecteren van kleine tot middelgrote interactie-effecten beperkt, wat in het bijzonder de interpretatie van het randsignificante attributie-effect op organisatieniveau bemoeilijkt. Bovendien kwam de steekproef tot stand via vrijwillige deelname langs sociale media en het persoonlijke netwerk van de onderzoeker, wat de representativiteit beperkt: de leeftijdsverdeling vertoont twee concentraties en hoger opgeleiden zijn oververtegenwoordigd.

Een tweede beperking betreft de vastgestelde baselineverschillen tussen de condities. Hoewel de randomisatiecontrole op demografische kenmerken geen onevenwicht aantoonde, verschilden de condities reeds op T1 in vertrouwensniveau. Dit onderzoek heeft daarmee rekening gehouden door de interpretatie te enten op de binnen-conditie veranderingen en de interacties met tijd, maar de between-subjects vergelijkingen blijven hierdoor slechts indicatief. Een grotere steekproef of een design met stratificatie op voorafgaand overheidsvertrouwen zou dit risico verkleinen.

Een derde beperking is inherent aan het vignetdesign. Fictieve scenario's over een fictieve gemeente meten intenties en percepties in een gecontroleerde setting, geen gedrag in een reële crisis. De externe validiteit is bovendien beperkt tot een type organisatie, een dienst sociale zaken, en een type incident. Ook werd het herstel onmiddellijk na de verontschuldiging gemeten; over de duurzaamheid van het effect doet dit cross-sectionele design geen uitspraken. Ten slotte kan bij de herstelratio's boven de honderd procent niet volledig worden uitgesloten dat regressie naar het gemiddelde of sociale wenselijkheid een rol speelde, al maakt het systematische verschil tussen de attributiecondities een louter artefactuele verklaring onwaarschijnlijk.

Een verdere beperking betreft de omvang en de verdeling van de uitval. Van de 140 beëindigde vragenlijsten bleven na toepassing van de manipulatiechecks en verwijdering van één onbruikbare observatie 93

respondenten over, een uitval van 34 procent. Die uitval was bovendien ongelijk verdeeld over de condities: in de twee competentiecondities viel respectievelijk 15,4 en 23,5 procent van de respondenten af, tegenover 46,7 en 51,4 procent in de twee integriteitscondities. De manipulatiecheck bij het integriteitsfalen bleek met andere woorden aanzienlijk moeilijker correct te beantwoorden dan die bij het competentiefalen, wat erop wijst dat het morele karakter van dat falen door een substantieel deel van de respondenten niet werd opgepikt zoals bedoeld. Dit heeft drie gevolgen.

Ten eerste verkleint het de celgroottes van de integriteitscondities tot 16 en 18 respondenten, wat de power voor precies die vergelijkingen ondermijnt. Ten tweede is uitsluiting op basis van manipulatiechecks geen willekeurig proces: wie de manipulatie niet correct waarnam, verschilt mogelijk systematisch van wie dat wel deed, bijvoorbeeld in aandacht of opleidingsniveau. Wanneer die uitval bovendien geconcentreerd is in twee van de vier condities, kan de oorspronkelijke randomisatie binnen de resterende groep gedeeltelijk verloren gaan, wat een plausibele aanvullende verklaring vormt voor de vastgestelde baselineverschillen op T1. Ten derde, en het meest ingrijpend, raakt deze selectiviteit rechtstreeks aan de toetsing van H1 en H3: de respondenten die in de integriteitscondities overbleven, zijn per definitie diegenen die het morele karakter van het falen wél herkenden, wat de vergelijking met de competentiecondities vertekent. Het uitblijven van elk effect van het type falen moet daarom met bijzondere terughoudendheid worden geïnterpreteerd: het is niet uit te sluiten dat de integriteitsmanipulatie eerder te zwak dan werkelijk irrelevant was.

Ten slotte verdient de gekozen manipulatie een kanttekening. Beide attributiecondities bevatten een volwaardige verontschuldiging met erkenning van het incident en een uiting van spijt; enkel de toegewezen oorzaak verschilde. Dit is een bewuste en theoretisch verantwoorde keuze, maar het betekent ook dat de gevonden effecten betrekking hebben op een relatief subtiel onderscheid binnen eenzelfde herstelstrategie. Sterkere contrasten, bijvoorbeeld tussen erkennen en ontkennen of tussen wel en niet communiceren, zouden mogelijk andere en grotere effecten opleveren.

5.6 Aanbevelingen voor toekomstig onderzoek

Uit de bevindingen en beperkingen vloeien verschillende onderzoekspistes voort. Longitudinaal onderzoek zou moeten nagaan of het vastgestelde herstel, en in het bijzonder het opvallende overherstel bij interne attributie, standhoudt over de tijd dan wel een tijdelijk effect van de communicatie is. Replicatie met grotere en meer representatieve steekproeven is nodig om uitsluitsel te geven over het onbesliste attributie-effect op organisatieniveau, en replicatie over meerdere typen publieke organisaties, zoals politie, zorg of federale agentschappen, zou de generaliseerbaarheid versterken.

Inhoudelijk verdient een studie die zowel de keuze tussen verontschuldiging en ontkenning (Kim et al., 2004) als de attributie binnen een verontschuldiging (Kim et al., 2006) in één publiek design combineert prioriteit, om na te gaan of het uitblijven van het matching-effect in dit onderzoek toe te schrijven is aan de publieke AI-context dan wel aan kenmerken van de steekproef of het vignet.

Daarnaast verdienen twee methodologische pistes aandacht. Een eerste is het meten van de mediators die het effect van attributie op herstel verklaren, zoals waargenomen oprechtheid, toegeschreven schuld en sympathie. Een mediatieanalyse zou kunnen aantonen of het voordeel van interne attributie inderdaad verloopt via de perceptie van eerlijkheid, zoals dit onderzoek op theoretische gronden veronderstelt. Een tweede piste is het afzonderlijk analyseren van de vertrouwensdimensies vermogen, welwillendheid en integriteit met meer items per dimensie. Op die manier kan empirisch worden nagegaan of de verontschuldiging inderdaad selectief op de integriteits- en welwillendheidsdimensie aangrijpt, wat de voorgestelde verklaring voor het verschil tussen organisatie- en leidersvertrouwen zou onderbouwen of weerleggen.

Ten slotte zou toekomstig onderzoek de rol van de leidinggevende verder kunnen ontleden. Dit onderzoek stelde vast dat een interne attributie het vertrouwen in de leidinggevende boven het oorspronkelijke niveau tilt, maar liet in het midden of dit effect voortkomt uit de inhoud van de boodschap, uit het feit dat een identificeerbare persoon haar uitsprekt, of uit de combinatie van beide. Een design waarin de boodschapper systematisch wordt gevarieerd, bijvoorbeeld tussen een benoemde leidinggevende en een onpersoonlijke woordvoerder van de dienst, zou hierover uitsluitsel kunnen geven en is rechtstreeks relevant voor de communicatiepraktijk van publieke organisaties.

5.7 Conclusie

De centrale onderzoeksvraag luidde in welke mate het type falen een modererend effect heeft op de relatie tussen de attributie in een verontschuldiging en het herstel van burgersvertrouwen in een publieke organisatie en haar leidinggevende. Het antwoord daarop is ondubbelzinnig negatief. Het veronderstelde moderatie-effect werd niet teruggevonden, noch voor het vertrouwen in de organisatie, noch voor het vertrouwen in de leidinggevende.

De drie deelvragen laten zich op basis van de resultaten als volgt beantwoorden. Op de eerste deelvraag, of een integriteitsfout tot een sterkere vertrouwensdaling leidt dan een competentiefout, luidt het antwoord ontkennend. De daling bedroeg in alle condities ongeveer een schaalpunt en verschilde niet significant naargelang het type falen. Burgers bleken in de eerste plaats te reageren op het feit dat een kwetsbare burger door een overheidsdienst werd benadeeld, en veel minder op de aard van de onderliggende fout. Op de tweede deelvraag, of een interne attributie effectiever is dan een externe, luidt het antwoord bevestigend maar voorwaardelijk. Het effect was duidelijk en robuust aanwezig voor het vertrouwen in de leidinggevende, terwijl het voor het organisatievertrouwen de conventionele significantiedrempel net niet haalde en gezien de beperkte power onbeslist blijft. Op de derde deelvraag, of het effect van de attributie verschilt naargelang het type falen, luidt het antwoord opnieuw ontkennend.

Daarmee levert dit onderzoek een bevinding op die genuanceerder is dan de hypothesen vooropstelden, maar die inhoudelijk niet minder betekenisvol is. Wat het herstel van vertrouwen bepaalt, is niet zozeer de

aard van de fout als wel de wijze waarop erover wordt gecommuniceerd, en dan vooral door wie. Een leidinggevende die de verantwoordelijkheid persoonlijk opneemt, herstelt niet alleen het geschonden vertrouwen maar overstijgt het oorspronkelijke niveau. Wie de oorzaak daarentegen buiten zichzelf legt, blijft achter met een structureel beschadigde vertrouwensrelatie. De verontschuldiging werkt met andere woorden het sterkst voor wie haar uitspreekt, en het minst voor wie zich erachter verschuilt.

Die vaststelling krijgt bijzondere betekenis in de context waarin dit onderzoek zich afspeelt. Naarmate publieke organisaties beslissingen aan algoritmen toevertrouwen, ontstaat een nieuwe en aantrekkelijke externe attributie: het systeem heeft gefaald, niet de organisatie. De resultaten suggereren dat die redenering communicatief een doodlopende weg is. Het algoritme is immers geen externe omstandigheid maar een instrument dat de organisatie zelf heeft gekozen, aangekocht en ingezet. Voor bestuurders die met de gevolgen van een AI-falen worden geconfronteerd, is de les van dit onderzoek dan ook concreet: vertrouwen wordt niet hersteld door de oorzaak aan te wijzen, maar door de verantwoordelijkheid op te nemen..

Literatuurlijst

- Alon-Barkat, S., & Busuioc, M. (2023). Human–AI interactions in public sector decision making: “Automation bias” and “selective adherence” to algorithmic advice. *Journal of Public Administration Research and Theory*, 33(1), 153–169. <https://doi.org/10.1093/jopart/muac007>
- Alon-Barkat, S., Busuioc, M., Schwoerer, K., & Weißmüller, K. S. (2025). Algorithmic discrimination in public service provision: Understanding citizens’ attribution of responsibility for human versus algorithmic discriminatory outcomes. *Journal of Public Administration Research and Theory*, 35(4), 469–488. <https://doi.org/10.1093/jopart/muaf024>
- Amnesty International. (2021). *Xenophobic machines: Discrimination through unregulated use of algorithms in the Dutch childcare benefits scandal*. Amnesty International.
- Aoki, N. (2020). An experimental study of public trust in AI chatbots in the public sector. *Government Information Quarterly*, 37(4), 101490. <https://doi.org/10.1016/j.giq.2020.101490>
- Bovens, M., & Zouridis, S. (2002). From street-level to system-level bureaucracies: How information and communication technology is transforming administrative discretion and constitutional control. *Public Administration Review*, 62(2), 174–184. <https://doi.org/10.1111/0033-3352.00168>
- Busuioc, M. (2021). Accountable artificial intelligence: Holding algorithms to account. *Public Administration Review*, 81(5), 825–836. <https://doi.org/10.1111/puar.13293>
- Coombs, W. T. (2007). Protecting organization reputations during a crisis: The development and application of Situational Crisis Communication Theory. *Corporate Reputation Review*, 10(3), 163–176. <https://doi.org/10.1057/palgrave.crr.1550049>
- de Bruijn, H., Warnier, M., & Janssen, M. (2022). The perils and pitfalls of explainable AI: Strategies for explaining algorithmic decision-making. *Government Information Quarterly*, 39(2), 101666. <https://doi.org/10.1016/j.giq.2021.101666>
- Dirks, K. T., Lewicki, R. J., & Zaheer, A. (2009). Repairing relationships within and between organizations: Building a conceptual foundation. *Academy of Management Review*, 34(1), 68–84. <https://doi.org/10.5465/amr.2009.35713285>
- Europese Commissie. (2024). *Verordening (EU) 2024/1689 tot vaststelling van geharmoniseerde regels betreffende artificiële intelligentie (AI Act)*. Publicatieblad van de Europese Unie.
- Glikson, E., & Woolley, A. W. (2020). Human trust in artificial intelligence: Review of empirical research. *Academy of Management Annals*, 14(2), 627–660. <https://doi.org/10.5465/annals.2018.0057>
- Grimmelikhuijsen, S. (2023). Explaining why the computer says no: Algorithmic transparency affects the perceived trustworthiness of automated decision-making. *Public Administration Review*, 83(2), 241–262. <https://doi.org/10.1111/puar.13483>
- Grimmelikhuijsen, S., & Knies, E. (2017). Validating a scale for citizen trust in government organizations. *International Review of Administrative Sciences*, 83(3), 583–601. <https://doi.org/10.1177/0020852315585950>
- Grimmelikhuijsen, S. G., Porumbescu, G., Hong, B., & Im, T. (2013). The effect of transparency on trust in government: A cross-national comparative experiment. *Public Administration Review*, 73(4), 575–586. <https://doi.org/10.1111/puar.12047>

- Janssen, M., & van den Hoven, J. (2015). Big and open linked data (BOLD) in government: A challenge to transparency and privacy? *Government Information Quarterly*, 32(4), 363–368. <https://doi.org/10.1016/j.giq.2015.11.007>
- Kim, P. H., Dirks, K. T., Cooper, C. D., & Ferrin, D. L. (2006). When more blame is better than less: The implications of internal vs. external attributions for the repair of trust after a competence- vs. integrity-based trust violation. *Organizational Behavior and Human Decision Processes*, 99(1), 49–65. <https://doi.org/10.1016/j.obhdp.2005.07.002>
- Kim, P. H., Ferrin, D. L., Cooper, C. D., & Dirks, K. T. (2004). Removing the shadow of suspicion: The effects of apology versus denial for repairing competence- versus integrity-based trust violations. *Journal of Applied Psychology*, 89(1), 104–118. <https://doi.org/10.1037/0021-9010.89.1.104>
- Kim, T., & Song, H. (2021). How should intelligent agents apologize to restore trust? Interaction effects between anthropomorphism and apology attribution on trust repair. *Telematics and Informatics*, 61, 101595. <https://doi.org/10.1016/j.tele.2021.101595>
- Lewicki, R. J., & Bunker, B. B. (1996). Developing and maintaining trust in work relationships. In R. M. Kramer & T. R. Tyler (Eds.), *Trust in organizations: Frontiers of theory and research* (pp. 114–139). Sage.
- Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An integrative model of organizational trust. *Academy of Management Review*, 20(3), 709–734. <https://doi.org/10.2307/258792>
- Meijer, A., & Grimmelikhuijsen, S. (2020). Responsible and accountable algorithmization: How to generate citizen trust in governmental usage of algorithms. In M. Schuilenburg & R. Peeters (Eds.), *The algorithmic society: Technology, power, and knowledge* (pp. 53–66). Routledge.
- OECD. (2022). *Building trust to reinforce democracy: Main findings from the 2021 OECD survey on drivers of trust in public institutions*. OECD Publishing. <https://doi.org/10.1787/b407f99c-en>
- Pareek, S., Velloso, E., & Goncalves, J. (2024). Trust development and repair in AI-assisted decision-making during complementary expertise. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24)*. ACM. <https://doi.org/10.1145/3630106.3658924>
- Reeder, G. D., & Brewer, M. B. (1979). A schematic model of dispositional attribution in interpersonal perception. *Psychological Review*, 86(1), 61–79. <https://doi.org/10.1037/0033-295X.86.1.61>
- Schiff, D. S., Schiff, K. J., & Pierson, P. (2022). Assessing public value failure in government adoption of artificial intelligence. *Public Administration*, 100(3), 653–673. <https://doi.org/10.1111/padm.12742>
- Selten, F., Robeer, M., & Grimmelikhuijsen, S. (2023). ‘Just like I thought’: Street-level bureaucrats trust AI recommendations if they confirm their professional judgment. *Public Administration Review*, 83(2), 263–278. <https://doi.org/10.1111/puar.13602>
- Tomlinson, E. C., Dineen, B. R., & Lewicki, R. J. (2004). The road to reconciliation: Antecedents of victim willingness to reconcile following a broken promise. *Journal of Management*, 30(2), 165–187. <https://doi.org/10.1016/j.jm.2003.01.003>
- van Bekkum, M., & Zuiderveen Borgesius, F. (2021). Digital welfare fraud detection and the Dutch SyRI judgment. *European Journal of Social Security*, 23(4), 323–340. <https://doi.org/10.1177/13882627211031257>

Van de Walle, S., & Bouckaert, G. (2003). Public service performance and trust in government: The problem of causality. *International Journal of Public Administration*, 26(8–9), 891–913. <https://doi.org/10.1081/PAD-120019352>

Van Thiel, S. (2021). *Bestuurskundig onderzoek: Een methodologische inleiding* (4e druk). Coutinho.

Weiner, B. (1986). *An attributional theory of motivation and emotion*. Springer. <https://doi.org/10.1007/978-1-4612-4948-1>

Wirtz, B. W., Weyerer, J. C., & Geyer, C. (2019). Artificial intelligence and the public sector — Applications and challenges. *International Journal of Public Administration*, 42(7), 596–615. <https://doi.org/10.1080/01900692.2018.1498103>

Young, M. M., Bullock, J. B., & Lecy, J. D. (2019). Artificial discretion as a tool of governance: A framework for understanding the impact of artificial intelligence on public administration. *Perspectives on Public Management and Governance*, 2(4), 301–313. <https://doi.org/10.1093/ppmgov/gvz014>