Hoe Apple Intelligence een nieuwe norm stelt voor AI-beveiliging, privacy en veiligheid

- E-mail

Iedereen had de aankondiging van Apple Intelligence verwacht, maar de details over beveiliging, privacy en veiligheid kwamen toch nog als een (welkome) verrassing voor een beveiligingsgemeenschap die al gewend is aan de sterke basisprincipes van Apple. Apple is er behendig in geslaagd om een reeks uitdagingen het hoofd te bieden met innovaties die zich uitstrekken van de iPhone tot de cloud en die alles overtreffen wat we elders hebben gezien. Ik werk al meer dan tien jaar in cloudbeveiliging en nog langer in cyberbeveiliging in het algemeen, en ik ben diep onder de indruk. Het ontwerp van Apple Intelligence en het Private Cloud Compute-systeem dat het ondersteunt, stelt een nieuwe norm voor de AI-industrie door gebruik te maken van een reeks investeringen die Apple de afgelopen tien jaar heeft gedaan op het gebied van beveiliging en privacy.
Om te begrijpen waarom dit zo belangrijk is en hoe Apple dit voor elkaar heeft gekregen (ervan uitgaande dat alles werkt zoals gedocumenteerd), moeten we beginnen met een kort overzicht van "dit" soort AI, de risico's die het met zich meebrengt en hoe Apple van plan is deze risico's aan te pakken.

Wat voor soort AI is Apple Intelligence en waarin verschilt het van Siri?

Er zijn veel soorten kunstmatige intelligentie, die allemaal wiskundige modellen gebruiken om problemen op te lossen op basis van leren, zoals het herkennen van patronen (alsjeblieft, AI-onderzoekers, doe me geen pijn voor deze simplificatie). Je iPhones en Macs vertrouwen al op AI voor tal van functies zoals Siri-spraakherkenning, het herkennen van gezichten in Foto's en beeldverbetering voor iPhone-foto's. In het verleden beschreef Apple deze functies als aangedreven door "machine learning", maar het bedrijf noemt ze nu "AI".
Inmiddels heb je ongetwijfeld gehoord van ChatGPT, een *generatieve AI-chatbot*. Generatieve AI-algoritmes creëren op verzoek nieuwe inhoud, waaronder tekst, afbeeldingen en meer. In tegenstelling tot Siri, die grotendeels beperkt is tot ingeblikte antwoorden, kan een generatieve AI-chatbot een veel breder scala aan verzoeken aan. Vraag Siri om je een verhaal te vertellen en het kan er een uit een database halen. Vraag ChatGPT om je een verhaal te vertellen en hij schrijft ter plekke een nieuw verhaal.
Open AI's ChatGPT, Anthropic's Claude en Google Gemini zijn allemaal voorbeelden van een *Generative Pre-trained Transformer* (GPT, snap je?). GPT's zijn in feite een subset van *Large Language Models* (LLM's) die vooraf zijn getraind op grote hoeveelheden tekst (grote stukken van het internet), die die gegevens gebruiken om nieuwe antwoorden te maken door te voorspellen welke woorden als volgende moeten komen als antwoord op een vraag. LLM's zijn voor tekst, maar andere vormen van generatieve AI creëren afbeeldingen, audio en zelfs video (die allemaal misbruikt kunnen worden voor deepfakes). Generatieve AI is zeer indrukwekkend, maar vereist enorme rekenkracht en mislukt vaak op spectaculaire wijze. Het creëert ook nieuwe beveiligingsproblemen en privacyproblemen en heeft te kampen met inherente veiligheidsproblemen.
Apple's eerste stap in generatieve AI wordt gezet onder de paraplu van Apple Intelligence. Apple geeft prioriteit aan beveiliging, privacy en veiligheid op manieren die bij eerdere AI-functies niet nodig waren.

Wat zijn de verschillen tussen AI-beveiliging, privacy en veiligheid?

Hoewel deze termen met elkaar te maken hebben en tot op zekere hoogte van elkaar afhankelijk zijn, hebben ze elk hun eigen domein:

- Bij**beveiliging** gaat het erom te voorkomen dat een tegenstander iets doet met het AI-systeem wat hij niet zou moeten doen. Een aanval die bekend staat als *prompt injection* bijvoorbeeld, probeert het model te verleiden om iets ongepasts te onthullen of te doen, zoals het onthullen van privégegevens van een andere gebruiker.
- **Privacy** zorgt ervoor dat je gegevens onder jouw controle blijven en door niemand zonder jouw toestemming kunnen worden bekeken of gebruikt, ook niet door de AI-provider. Je vragen aan de AI moeten privé blijven en onleesbaar voor anderen.
- **Veiligheid** betekent dat de AI nooit schadelijke antwoorden mag geven of schadelijke acties mag ondernemen. Een AI mag je niet vertellen hoe je jezelf kunt verwonden, een biologisch wapen kunt maken of een bank kunt beroven.

Beveiliging is de basis waarop privacy en veiligheid zijn gebouwd; als het systeem onveilig is, kunnen we geen privacy of veiligheid garanderen.
Zoals bij elke online dienst is privacy een keuze; aanbieders kiezen welke privacyopties ze aanbieden en consumenten kiezen of en hoe ze een dienst gebruiken. Veel AI-aanbieders voor consumenten gebruiken standaard je aanwijzingen (de vragen die je aan de AI stelt) om hun modellen te verbeteren. Dit betekent dat alles wat je invoert, waarschijnlijk fragmentarisch, kan worden gebruikt in het antwoord van iemand anders. Aan de andere kant kun je er bij de meeste aanbieders voor kiezen om je prompts niet te laten gebruiken voor training en bieden ze opties om je gegevens en geschiedenis te verwijderen.
De meeste AI-aanbieders doen hun best om de veiligheid te garanderen, maar net als sociale netwerken gebruiken ze verschillende definities en hebben ze verschillende toleranties voor wat ze acceptabel vinden. Het is onvermijdelijk dat sommige mensen het daar niet mee eens zijn.

Hoe werkt generatieve AI?

AI is ongelooflijk complex, maar voor ons doel kunnen we het vereenvoudigen en ons richten op drie kerncomponenten en een paar extra opties. Deze worden gecombineerd om een *model* te produceren:

- **Hardware om het AI-model uit te voeren:** Hoewel modellen op gewone CPU's kunnen draaien, hebben ze baat bij gespecialiseerde chips die ontworpen zijn om speciale soorten software te draaien die veel voorkomen in AI. Dit zijn meestal GPGPU's (General-Purpose Graphics Processing Units) afgeleid van grafische chips, voornamelijk van beurslieveling Nvidia, of speciale AI-chips zoals Apple's Neural Engine en Google's Tensor Processing Units.
- **AI-software/algoritmen:** Dit is het brein van de modellen en bestaat uit meerdere componenten. De meeste huidige modellen maken gebruik van neurale netwerken, die nabootsen hoe een biologisch neuron (hersencel) werkt en communiceert met andere neuronen.
- **Trainingsgegevens:** Alle generatieve AI-modellen hebben een corpus aan kennis nodig om van te leren. De huidige consumentenmodellen zoals ChatGPT, Claude en Gemini zijn getraind door het web te schrapen, net zoals zoekmachines het web schrapen om indexen op te bouwen. Dit is controversieel en er lopen rechtszaken.

Zoals je je kunt voorstellen, levert een groter brein dat bestaat uit een groter aantal efficiënter verbonden neuronen die getraind zijn op een grotere dataset over het algemeen betere resultaten op. Wanneer een bedrijf een groot model bouwt dat ontworpen is voor algemeen gebruik, noemen we het een *foundation model*. Foundation-modellen kunnen in veel verschillende situaties worden geïntegreerd en verbeterd voor specifieke gebruikssituaties, zoals het schrijven van programmacode.
Twee belangrijke optionele componenten worden gebruikt om foundation modellen te verbeteren:

- **Fine-tuning data** past een vooraf getraind model aan om gepersonaliseerde resultaten te leveren. U kunt bijvoorbeeld een foundationmodel dat menselijke taal begrijpt verfijnen met voorbeelden van uw eigen handschrift om uw persoonlijke stijl na te bootsen. Het verfijnen van een model breidt de trainingsgegevens uit met meer specifieke gegevens.
- **Retrieval Augmented Generation (RAG)** verbetert de uitvoer door externe gegevens op te nemen die niet in de voorgetrainde of verfijnde gegevens zitten. Waarom RAG gebruiken in plaats van finetuning? Bij fine-tuning wordt het model opnieuw getraind, wat meer tijd en rekenkracht vergt. RAG is beter voor het verbeteren van resultaten met actuele gegevens. Afhankelijk van de grootte van het model en het volume aan fine-tuninggegevens, kan het weken of langer duren om een model bij te werken.

Dit is hoe het allemaal in elkaar steekt, aan de hand van het voorbeeld van de integratie van AI in een helpsysteem. De AI-ontwikkelaars van het basismodel produceren een nieuw LLM dat ze op een enorm rekencluster laden en vervolgens trainen met een enorme dataset. Het resultaat is zoiets als ChatGPT, dat de talen waarop het getraind is "begrijpt" en erin schrijft. Als reactie op een prompt beslist het welke woorden in welke volgorde moeten worden gezet, gebaseerd op al dat leren en de statistische waarschijnlijkheid van hoe verschillende woorden geassocieerd en met elkaar verbonden zijn.
Een klant past het basismodel vervolgens aan door zijn eigen verfijningsgegevens toe te voegen, zoals documentatie voor zijn softwareplatforms, en de LLM te integreren in zijn helpsysteem. Het basismodel begrijpt taal en de fijnafstemming levert specifieke details over die platforms. Als er veel verandert, kunnen de ontwikkelaars RAG gebruiken om het afgestemde model de nieuwste documentatie te laten ophalen en de resultaten te vergroten zonder opnieuw te hoeven trainen en afstemmen.
Als het werkt, voelt het als magie. Het probleem is dat LLM's (en andere generatieve AI, inclusief beeldgeneratoren) gevoelig zijn voor mislukkingen. Ze begrijpen statistisch waarschijnlijke associaties, maar niet noodzakelijkerwijs... de werkelijkheid. Ik heb ChatGPT eens gevraagd om een lijst met PG-13 komedies voor de familiefilmavond. De helft van de resultaten was R (inclusief degene die we kozen, dus misschien was het toch niet zo verkeerd). Dit soort fouten worden vaak "hallucinaties" genoemd en er wordt algemeen aangenomen dat ze nooit helemaal uit te sluiten zijn. Sommigen hebben gesuggereerd dat "confabulaties" misschien een betere term is, omdat "hallucinatie" connotaties heeft van wilde fantasie, terwijl "confabulatie" meer te maken heeft met verzinsels zonder de bedoeling om te misleiden.

Hoe beïnvloedt generatieve AI veiligheid, privacy en beveiliging?

Alleen al de complexiteit van generatieve AI creëert een breed scala aan nieuwe veiligheidsproblemen. In plaats van te proberen ze allemaal te behandelen, zullen we ons richten op hoe ze van invloed kunnen zijn op de AI-diensten die Apple levert aan iPhone-gebruikers. De kern van het probleem is dat Apple Intelligence alleen goed kan zijn als het minstens gedeeltelijk in de cloud draait en genoeg hardwaremogelijkheden heeft. Dit zijn enkele uitdagingen waar Apple voor staat:

- Om persoonlijke resultaten te kunnen bieden, hebben de AI-modellen toegang nodig tot persoonlijke gegevens die Apple liever niet verzamelt.
- Je kunt maar zoveel doen op één apparaat. Foundation-modellen draaien meestal in de cloud vanwege de enorme verwerkingsvereisten. Voor personalisatie moeten persoonlijke gegevens dus in de cloud worden verwerkt.
- Een risico van AI is dat een aanvaller een model kan misleiden om gegevens te onthullen die het niet zou moeten onthullen. Dat kunnen persoonlijke gebruikersgegevens zijn (zoals je prompts) of veiligheidsovertredingen (zoals informatie over de meest effectieve manier om een lichaam in de woestijn te begraven, hoewel dit hier in Phoenix als algemene kennis wordt beschouwd).
- Alles wat in de cloud draait, staat open voor aanvallen van buitenaf. Een beveiligingsincident in de cloud kan resulteren in een privacyschending waarbij klantgegevens worden onthuld.
- Apple is groot, populair en het doelwit van de meest geavanceerde cyberaanvallen die de mensheid kent. Kwaadwillenden en overheden zouden graag toegang hebben tot de persoonlijke vragen en e-mailoverzichten van een miljard gebruikers.

Deze uitdagingen zijn zeer complex. De meeste grote foundationmodellen zijn redelijk goed beveiligd, maar ze hebben toegang tot alle klantvragen. Voor Apple zijn de problemen nog groter omdat iPhones, iPads en Macs zo persoonlijk zijn en dus lokaal en in iCloud toegang hebben tot privégegevens. Niemand vraagt Apple om nog een generieke AI-chatbot te maken om ChatGPT te vervangen - mensen willen een Apple AI die hen begrijpt en persoonlijke resultaten geeft op basis van de informatie op hun iPhones en in iCloud.
De uitdaging voor Apple is om de kracht van generatieve AI op een veilige manier in te zetten, door gebruik te maken van de meest persoonlijke gegevens en deze gegevens privé te houden, zelfs voor intimi, criminelen en overheden.

Hoe gaat Apple Intelligence om met beveiliging en privacy?

De aanpak van Apple maakt gebruik van zijn volledige controle over de hardware- en softwarestacks op onze apparaten. Apple Intelligence probeert eerst een AI-aanvraag te verwerken op het lokale systeem (je iPhone, iPad of Mac) met behulp van Neural Engine-kernen die zijn ingebouwd in de A17 Pro- of M-serie chip. Als een taak meer verwerkingskracht vereist, stuurt Apple Intelligence het verzoek door naar een systeem genaamd *Private Cloud Compute*, dat de AI-modellen van Apple in de cloud uitvoert.
In plaats van te vertrouwen op publieke basismodellen, heeft Apple zijn eigen basismodellen gebouwd en draait deze op zijn eigen clouddienst, aangedreven door Apple siliciumchips, waarbij veel van dezelfde beveiligingsmogelijkheden worden gebruikt die onze persoonlijke Apple apparaten beschermen. Apple heeft deze mogelijkheden vervolgens uitgebreid met extra beveiligingen om ervoor te zorgen dat niemand toegang kan krijgen tot klantgegevens, inclusief kwaadwillende Apple medewerkers, mogelijke fabrieken in de fysieke of digitale

toeleveringsketen van Apple en overheidsspionnen.
Het is een verbazingwekkende daad van beveiliging en privacy engineering. Ik ben niet geneigd tot superlatieven - beveiliging is complex en er zijn *altijd* zwakke plekken die tegenstanders kunnen uitbuiten - maar dit is een van de weinige situaties in mijn carrière waarin ik denk dat superlatieven gerechtvaardigd zijn.
Hoewel Apple nog niet alle details heeft onthuld over hoe Apple Intelligence zal werken, heeft het bedrijf wel een overzicht gepubliceerd van de basismodellen, een overzicht van de beveiliging van Private Cloud Compute en informatie over de Apple Intelligence-functies. Door deze informatie te combineren met de eerder gepubliceerde Platform Security Guide van het bedrijf, waarin gedetailleerd wordt beschreven hoe Apple beveiliging toepast op apparaten en diensten, kunnen we de basisprincipes begrijpen van hoe Apple Apple Intelligence wil beveiligen.

Welke basismodellen gebruikt Apple en bevatten deze ook klantgegevens?

Apple heeft zijn basismodellen gemaakt met behulp van het Apple AXLearnframework, dat in 2023 als open-sourceproject is vrijgegeven. Houd er rekening mee dat een model het resultaat is van verschillende softwarealgoritmen die zijn getraind op een corpus van gegevens.
Apple gebruikt geen klantgegevens in de training, maar wel gelicentieerde gegevens en internetgegevens die zijn verzameld met een tool genaamd AppleBot, die het web afspeurt. Hoewel AppleBot niet nieuw is, hebben maar weinig mensen er tot nu toe veel aandacht aan besteed. Omdat persoonlijke gegevens van het internet opduiken in trainingsgegevens, probeert Apple dergelijke details eruit te filteren.
Net als andere makers van foundationmodellen heeft Apple enorme hoeveelheden tekst nodig om de capaciteiten van zijn modellen te trainen. Web scraping is omstreden omdat deze tools zonder toestemming intellectueel eigendom opgraven voor integratie in modellen en zoekindices. Ik heb ChatGPT ooit een vraag gesteld over cloudbeveiliging, een onderwerp waarover ik uitgebreid heb gepubliceerd, en het resultaat leek erg op wat ik in het verleden heb geschreven. Weet ik zeker dat het mij kopieerde? Nee, maar ik weet wel dat de crawler van ChatGPT mijn inhoud heeft gescraped.
Apple zegt nu dat het mogelijk is om je website uit te sluiten van AppleBot's crawling, maar alleen in de toekomst. Apple heeft niets gezegd over een manier om inhoud te verwijderen uit zijn bestaande foundation-modellen, die werden getraind voordat de uitsluitingsregels bekend werden. Dat ziet er niet goed uit voor het bedrijf, maar het zou waarschijnlijk vereisen dat het model opnieuw wordt getraind op de opgeschoonde dataset, wat zeker een mogelijkheid is.
Apple filtert ook op godslastering en inhoud van lage waarde; hoewel we het niet zeker weten, wordt schadelijke inhoud waarschijnlijk ook zoveel mogelijk uitgefilterd. Tijdens het trainen integreren de modellen ook menselijke feedback en voert Apple tests uit om beveiligingsproblemen (waaronder prompt injection) en veiligheidskwesties (zoals schadelijke inhoud) aan te pakken.

Is het basismodel afgestemd op mijn gegevens?

Apple verfijnt *adapters* in plaats van het basismodel. Adapters zijn verfijnde lagen voor verschillende taken, zoals het samenvatten van e-mailberichten. Om terug te komen op mijn algemene beschrijving van generatieve AI: Apple stemt een kleinere adapter af in plaats van het hele model - net zoals mijn voorbeeldbedrijf zijn helpsysteem voor productdocumentatie afstemt.
Apple schakelt dan een geschikte adapter in op basis van de taak die de gebruiker uitvoert. Dit lijkt een elegante manier om te optimaliseren voor zowel verschillende gebruikssituaties als de beperkte bronnen van een lokaal apparaat.
Op basis van de documentatie van Apple lijkt de fijnafstemming geen gebruik te maken van persoonlijke gegevens, vooral omdat de fijnafgestelde adapters worden getest en geoptimaliseerd voordat ze worden vrijgegeven, wat niet mogelijk zou zijn als ze werden getraind op individuele gegevens. Apple gebruikt ook verschillende basismodellen op het apparaat en in de cloud, en stuurt alleen de vereiste persoonlijke semantische gegevens naar de cloud voor elk verzoek.

Hoe werkt Apple Intelligence veilig op onze apparaten?

In sommige opzichten is het handhaven van de beveiliging op onze apparaten het makkelijkste deel van het probleem voor Apple, dankzij meer dan tien jaar werk aan het bouwen van veilige apparaten. Apple moet twee grote problemen op het apparaat oplossen:

- Gegevens verwerken en beschikbaar maken voor de AI-modellen
- Communiceren met de cloudservice wanneer dat nodig is

Zoals gezegd zal Apple Intelligence eerst kijken of het een verzoek on-device kan verwerken. Daarna wordt de juiste adapter geladen. Als de taak toegang tot je persoonlijke gegevens vereist, wordt dat op het apparaat zelf afgehandeld met behulp van een semantische index die lijkt op die van Spotlight. Apple heeft niet gespecificeerd hoe dit gebeurt, maar ik vermoed dat RAG wordt gebruikt om de benodigde gegevens uit de index te halen. Dit werk wordt gedaan met behulp van verschillende onderdelen van Apple silicium, met name de Neural Engine. Daarom werkt Apple Intelligence niet op alle apparaten: het heeft een voldoende krachtige Neural Engine en voldoende geheugen nodig.
Op dit punt is uitgebreide hardwarebeveiliging in het spel, veel verder dan wat ik in dit artikel kan behandelen. Apple maakt gebruik van meerdere lagen van encryptie, veilig geheugen en veilige communicatie op de chips uit de A- en M-serie om ervoor te zorgen dat alleen goedgekeurde applicaties met elkaar kunnen praten, dat gegevens veilig worden bewaard en dat geen enkel proces kan worden gecompromitteerd om het hele systeem te breken.
App-gegevens worden standaard niet geïndexeerd, zodat Apple je bankgegevens niet kan zien. Alle apps op iOS zijn gecompartimenteerd met verschillende coderingssleutels en de ontwikkelaar van een app moet zijn gegevens "publiceren" in de index. Deze gegevens omvatten *intenties*, dus een app kan niet alleen informatie publiceren, maar ook acties, die Apple Intelligence beschikbaar kan maken voor Siri. Ontwikkelaars kunnen ook semantische informatie voor hun apps publiceren (bijvoorbeeld definiëren wat een reisroute is).
Met andere woorden, je app-gegevens worden niet opgenomen in Apple Intelligence tenzij de ontwikkelaar het toestaat en ervoor bouwt. Ik vermoed ook dat gebruikers in staat zullen zijn om de opname uit te schakelen, aangezien die mogelijkheid al bestaat voor Siri en Spotlight in Instellingen > Siri & Zoeken op de iPhone en iPad en Systeeminstellingen > Siri & Spotlight > Siri Suggesties & Privacy op de Mac.
De bestaande beveiliging op het apparaat beperkt ook welke informatie een app kan zien, zelfs als een Siri-verzoek je persoonlijke gegevens combineert met app-gegevens. Siri zal alleen beschermde gegevens aan een app verstrekken als onderdeel van een Siri-verzoek als

die app al toegang heeft tot die beschermde gegevens (zoals wanneer je een berichten-app toegang geeft tot Contacten). Ik verwacht dat dit ook zo zal blijven voor Apple Intelligence, om iets te voorkomen dat beveiligingsprofessionals het *'confused deputy' probleem* noemen. Dit ontwerp moet voorkomen dat een kwaadwillende app het besturingssysteem misleidt om privégegevens van een andere app te verstrekken.

Hoe beslist Apple Intelligence of een verzoek naar de cloud moet gaan?

Apple heeft dit proces nog niet in detail beschreven, maar we kunnen wel wat conclusies trekken.
Zoals ik al zei, noemen we een verzoek aan de meeste vormen van generatieve AI een *prompt*, zoals "proeflees dit document". Eerst zet de AI de prompt om in *tokens*. Een token is een brok tekst die een AI gebruikt om te verwerken. Een maatstaf voor de kracht van een LLM is het aantal tokens dat het kan verwerken. De *woordenschat* van een model is alle tokens die het kan herkennen.
In mijn proefleesvoorbeeld hierboven is het aantal tokens gebaseerd op de grootte van het verzoek en de grootte van de gegevens (het document) die in het verzoek worden verstrekt. In het overzicht van het model zegt Apple dat de woordenschat 49.000 tokens is op apparaten en 100.000 in de cloud. Een iPhone 15 Pro kan antwoorden genereren met een snelheid van 30 tokens per seconde.
Ik denk dat Apple minstens twee criteria gebruikt om te beslissen of een verzoek naar de cloud wordt gestuurd:

- Vragen met meer dan een bepaald aantal tokens
- Vragen om bepaalde soorten informatie waarvoor een groter vocabulaire, meer geheugen of verwerkingskracht nodig is

Wat gebeurt er als een verzoek naar de cloud wordt gestuurd?

Hier heeft Apple zichzelf overtroffen met zijn beveiligingsmodel. Het bedrijf had een mechanisme nodig om de prompt veilig naar de cloud te sturen met behoud van de privacy van de gebruiker. Vervolgens moet het systeem die prompts verwerken - die gevoelige persoonlijke gegevens bevatten - zonder dat Apple of iemand anders toegang krijgt tot die gegevens. Tot slot moet het systeem de wereld verzekeren dat de vorige twee stappen verifieerbaar waar zijn. In plaats van ons simpelweg te vragen het te vertrouwen, heeft Apple meerdere mechanismen ingebouwd zodat je apparaat weet of het de cloud kan vertrouwen en de wereld weet of het Apple kan vertrouwen.
Hoe pakt Apple al deze uitdagingen aan? Met zijn nieuwe Private Cloud Compute: een op maat gemaakte hardware- en softwarestack voor het veilig en privé hosten van LLM's met verifieerbaar vertrouwen.
Uit de aankondiging van Private Cloud Compute:

We hebben Private Cloud Compute zo ontworpen dat het verschillende garanties biedt over de manier waarop het omgaat met gebruikersgegevens:

- *Het apparaat van een gebruiker stuurt gegevens naar PCC met als enige, exclusieve doel om te voldoen aan het inferentieverzoek van de gebruiker. PCC gebruikt die gegevens alleen om de door de gebruiker gevraagde bewerkingen uit te voeren.*
- *Gebruikersgegevens blijven alleen op de knooppunten van PCC die de aanvraag verwerken totdat het antwoord is teruggestuurd. PCC verwijdert de gegevens van de*

gebruiker nadat de aanvraag is uitgevoerd en er worden geen gebruikersgegevens in welke vorm dan ook bewaard nadat het antwoord is teruggestuurd.
- *Gebruikersgegevens zijn nooit beschikbaar voor Apple, zelfs niet voor personeel met administratieve toegang tot de productieservice of -hardware.*

Op hoog niveau valt Private Cloud Compute in een familie van mogelijkheden die we *vertrouwelijk computergebruik*noemen. Vertrouwelijke gegevensverwerking wijst specifieke hardware toe aan een bepaalde taak en die hardware is beveiligd tegen aanvallen of afluisteren door iemand met fysieke toegang. Het bedreigingsmodel van Apple omvat iemand met fysieke toegang tot de hardware en zeer geavanceerde vaardigheden - ongeveer het moeilijkste scenario om je tegen te verdedigen. Een ander voorbeeld is de Nitro-architectuur van Amazon Web Service.
Laten we Private Cloud Compute in hapklare brokken verdelen: het is behoorlijk complex, zelfs voor een levenslange beveiligingsprofessional met ervaring in cloud computing en vertrouwelijke gegevensverwerking zoals ik.

Welke gegevens worden naar de cloud gestuurd?

De prompt, het gewenste AI-model en eventuele ondersteunende gegevens. Ik denk dat dit ook contact- of app-gegevens zijn die niet zijn opgenomen in de prompt die de gebruiker typt of uitspreekt.

Hoe weet mijn apparaat waar het het verzoek naartoe moet sturen en hoe zorg ik ervoor dat het veilig en privé is?

De kerneenheid van Private Cloud Compute (PCC) is een *knooppunt*. Apple heeft niet gespecificeerd of een node een verzameling servers is of een verzameling processors op een enkele server, maar dat is vanuit beveiligingsoogpunt grotendeels irrelevant. PCC nodes gebruiken een niet nader gespecificeerde Apple silicium processor met dezelfde Secure Enclave als andere Apple apparaten. De Secure Enclave handelt encryptie af en beheert encryptiesleutels buiten de CPU. Zie het als een zeer veilige kluis, met een beetje verwerkingscapaciteit die alleen beschikbaar is voor beveiligingsoperaties.
Elk knooppunt heeft zijn eigen digitale certificaat, dat de publieke sleutel van het knooppunt en wat standaard metadata bevat, zoals wanneer het certificaat verloopt. De privésleutel die gekoppeld is aan de publieke sleutel wordt opgeslagen in de Secure Enclave op de server van het knooppunt. Gegevens die zijn versleuteld met een openbare sleutel kunnen alleen worden ontsleuteld met de bijbehorende privésleutel. Dit is publieke sleutel cryptografie, die in principe overal wordt gebruikt.
Clientsoftware op het apparaat van de gebruiker neemt eerst contact op met de loadbalancer van PCC met wat eenvoudige metadata, waardoor het verzoek kan worden doorgestuurd naar een geschikte node voor het benodigde model. De load balancer stuurt een lijst met knooppunten terug die klaar zijn om de aanvraag van de gebruiker te verwerken. Het apparaat van de gebruiker versleutelt vervolgens het verzoek met de publieke sleutels van de geselecteerde knooppunten, die nu de enige hardware zijn die de gegevens kunnen lezen. Vergeet niet dat dankzij de Secure Enclave er geen manier zou moeten zijn om de private sleutels van de nodes te achterhalen (een probleem met software-only encryptiesystemen), en dus zou er geen manier moeten zijn om het verzoek buiten die servers te lezen. De loadbalancer zelf kan de verzoeken niet lezen-het routeert ze gewoon naar de juiste nodes. Zelfs als een aanvaller de loadbalancer zou compromitteren en het verkeer naar andere hardware zou sturen, zou die hardware het verzoek nog steeds niet kunnen lezen omdat hij de ontcijferingssleutels niet heeft.

Kan iemand nagaan welk verzoek van mij is of mijn verzoek naar een specifieke node sturen?

Nee, en dit is een erg coole functie. Kortom, Apple kan je IP-adres of apparaatgegevens niet zien omdat het een relay van een derde partij gebruikt die dergelijke informatie verwijdert. Die derde partij kan zich echter ook niet voordoen als Apple of gegevens ontsleutelen.
De initiële metadata die naar de load balancer wordt gestuurd om de lijst met nodes op te vragen, bevat geen identificerende informatie. Het zegt in wezen: "Ik heb een model nodig voor het proeflezen van mijn document". Dit verzoek gaat niet rechtstreeks naar Apple, maar wordt doorgestuurd naar een derde partij die het IP-adres en andere identificerende informatie verwijdert.
Apple kan een verzoek dus niet terugleiden naar een apparaat, waardoor een aanvaller niet hetzelfde kan doen tenzij hij zowel Apple als de relay service kan compromitteren. Mocht een aanvaller daadwerkelijk een knooppunt compromitteren en er een specifiek doel naartoe willen sturen, dan verdedigt Apple zich verder tegen sturing door statistische analyses van loadbalancers uit te voeren om eventuele onregelmatigheden te detecteren in waar verzoeken naartoe worden gestuurd.
Tot slot zegt Apple hier niets over in zijn documentatie, maar we kunnen hieruit afleiden dat de knooppuntcertificaten zijn ondertekend met de speciale ondertekensleutels die zijn ingebed in de besturingssystemen en hardware van Apple. Apple beschermt deze als kroonjuwelen. Deze handtekeningverificatie voorkomt dat een aanvaller zich voordoet als een officieel Apple knooppunt. Je apparaat versleutelt een verzoek voor de nodes die zijn opgegeven door de load balancer, zodat zelfs andere PCC-nodes je verzoek niet kunnen lezen.

Hoe worden mijn gegevens beschermd in Apple Intelligence dat in de cloud draait?

Op dit punt in het proces heeft je apparaat gezegd: "Ik heb PCC nodig voor een proefleesverzoek", en de relay service van Apple heeft geantwoord: "Hier is een lijst met nodes die dat kunnen leveren". Vervolgens controleert je apparaat de certificaten en sleutels voordat het de aanvraag versleutelt en naar de nodes stuurt.
De load balancer stuurt je verzoek vervolgens door naar de nodes. Vergeet niet dat de nodes draaien op speciale Apple servers die speciaal voor PCC zijn gebouwd. Deze servers maken gebruik van dezelfde beproefde beveiligingsmechanismen als je persoonlijke Apple apparaten, maar dan verder versterkt om bescherming te bieden tegen geavanceerde aanvallen. Hoe?
Alle hardware van PCC wordt gebouwd in een beveiligde toeleveringsketen en elke server wordt grondig geïnspecteerd voordat hij klaar is voor gebruik. (Deze technieken zijn essentieel om te voorkomen dat er achterdeurtjes worden ingebouwd voordat de servers Apple zelfs maar bereiken). Het algehele proces lijkt beeldvorming met hoge resolutie, testen en het volgen van de toeleveringsketen te omvatten, en alles wordt door een derde partij gecontroleerd. Zodra een server deze controles doorstaat, wordt hij verzegeld en wordt een sabotageschakelaar geactiveerd om elke poging tot fysieke wijziging te detecteren.
Wanneer een server wordt opgestart, wordt het Secure Boot-proces van Apple gebruikt, dat wordt beschreven in de Platform beveiligingsgids. Deze aanpak maakt gebruik van de Secure Enclave en meerdere stappen om ervoor te zorgen dat het geladen besturingssysteem geldig is, code niet kan worden geïnjecteerd en alleen goedgekeurde applicaties kunnen worden uitgevoerd. Net als macOS gebruiken PCC-servers een Signed System Volume, wat betekent dat het besturingssysteem cryptografisch is ondertekend om aan te tonen dat er niet mee is geknoeid en draait op alleen-lezen opslag.
Alle software die op PCC-servers draait, is ontwikkeld en ondertekend door Apple, waardoor de kans op problemen door een kwaadwillende ontwikkelaar die een open source tool compromitteert, wordt verkleind. Veel software is geschreven in Swift, een geheugenveilige taal die bestand is tegen het kraken van bepaalde exploits. En alles maakt gebruik van

sandboxing en andere standaard Apple software beveiligingscontroles, net als je iPhone.
Bij het opstarten worden willekeurige versleutelingscodes gegenereerd voor het datavolume (de opslag die wordt gebruikt voor het verwerken van verzoeken). Je gegevens worden dus versleuteld opgeslagen op de server en alles wordt beveiligd met de Secure Enclave.
Het wordt nog beter. Nadat een node een verzoek heeft verwerkt, gooit Apple de versleutelingscodes weg en start de node opnieuw op. Dat knooppunt kan geen eerder opgeslagen gebruikersgegevens meer lezen omdat het niet langer over de coderingssleutel beschikt! Het hele systeem reset zichzelf voor de volgende aanvraag. Voor de zekerheid recycleert Apple zelfs af en toe het geheugen van de server voor het geval daar nog iets opgeslagen was.
Om het in elkaar te passen: nadat je je verzoek naar Apple hebt gestuurd, gaat het naar de streng beveiligde Private Cloud Compute nodes. Die verwerken het verzoek, waarbij je gegevens de hele tijd versleuteld blijven. Als de aanvraag is voltooid, wissen ze zichzelf cryptografisch, starten ze opnieuw op en zijn ze klaar voor de volgende aanvraag.

Worden mijn gegevens blootgesteld aan beheerders, in logbestanden of aan tools voor prestatiebewaking?

Nee. Apple bevat geen software die dit soort monitoring (*privileged runtime access* genaamd) in de stack mogelijk maakt. PCC nodes hebben geen command shells, debugging modi of ontwikkelaarstools. Prestatie- en loggingtools zijn beperkt en ontworpen om alle privégegevens te verwijderen.

Hoe gaat Apple bewijzen dat we nodes kunnen vertrouwen... en zijn woord?

Apple zegt dat het elke productiesoftware-build van Private Cloud Compute publiekelijk beschikbaar zal maken voor onderzoekers om te evalueren. Apparaten zullen alleen verzoeken sturen naar nodes die kunnen aantonen dat ze een van deze openbare builds draaien. Dit is een ander uniek onderdeel van het Apple Intelligence ecosysteem.
Apple zal dit bereiken door gebruik te maken van een openbaar *transparantielogboek*, dat cryptografie gebruikt om ervoor te zorgen dat als er eenmaal iets naar het logboek is geschreven, het niet meer kan worden veranderd - een goed gebruik van blockchaintechnologie. Dit logboek zal metingen van de code bevatten (momenteel niet gespecificeerd) die kunnen worden gebruikt om te valideren dat een binaire blob van het besturingssysteem en de applicaties overeenkomt met de gelogde versie.
Bovendien zal Apple de binaire afbeeldingen publiceren van de softwarestack die op PCC-nodes draait. Dat geeft vertrouwen en is een geweldige manier om ervoor te zorgen dat het systeem echt veilig is, en niet alleen "veilig" omdat het obscuur is.
Zoals gezegd sturen onze apparaten alleen verzoeken naar nodes waarop verwachte software-images draaien. Apple is hier een beetje vaag over, maar ik vermoed dat de nodes ook hun cryptografisch ondertekende metingen zullen publiceren, die moeten overeenkomen met de metingen voor de huidige versie van de software die gepubliceerd is in het transparantielogboek.
Tot slot zal Apple ook een PCC Virtual Research Environment vrijgeven voor beveiligingsonderzoekers om hun claims te valideren door het gedrag van een lokale, virtuele versie van een PCC-node te testen. Het bedrijf zal ook wat broncode vrijgeven, waaronder wat plain-text code voor gevoelige onderdelen die het bedrijf nog niet eerder heeft vrijgegeven.

Maar stuurt Apple geen verzoeken naar ChatGPT?

Apple Intelligence richt zich op AI-taken die draaien om je apparaten en gegevens. Voor meer algemene verzoeken die wat Apple wereldkennis noemt vereisen, zal Apple Intelligence de gebruiker vragen het verzoek te sturen - in eerste instantie naar ChatGPT en in de toekomst naar andere diensten.
Elke prompt vereist toestemming van de gebruiker (wat vervelend kan worden), en Apple heeft verklaard dat OpenAI (dat ChatGPT uitvoert) geen IP-adressen zal traceren voor anonieme verzoeken. Als je een betaalde account hebt bij ChatGPT of een andere AI-dienst van derden die Apple in de toekomst zal ondersteunen, zal de privacy door die dienst worden afgehandeld volgens zijn privacybeleid.

Hoe zit het met China en andere landen waar Apple verplicht is om de overheid toegang te geven tot gegevens? Hoe zit het met de Europese Unie?

Apple heeft niet gezegd of het Apple Intelligence zal vrijgeven in landen als China. Het gedocumenteerde Private Cloud Compute-model voldoet niet aan de Chinese wetgeving. Apple heeft al aangekondigd dat het Apple Intelligence in eerste instantie niet zal vrijgeven in de EU vanwege de Digital Markets Act. Hoewel Apple Intelligence uiteindelijk in staat zal zijn om verzoeken naar diensten van derden te sturen voor wereldkennis, omvatten deze verzoeken niet de privégegevens die op het apparaat of in PCC worden verwerkt. Het is moeilijk in te zien hoe Apple de privacy van gebruikers kan handhaven en tegelijkertijd een externe dienst dezelfde diepgaande toegang tot gegevens op het apparaat kan geven, die de EU zou kunnen eisen voor naleving van de DMA.

Waarom heeft Apple zoveel moeite gedaan om een eigen AI-platform te bouwen?

Hoewel veel mensen kunstmatige intelligentie liever afdoen als de nieuwste technologische rage, is de kans groot dat het in de loop der tijd een grote invloed op ons leven en onze maatschappij zal hebben.
AI-modellen blijven zich razendsnel ontwikkelen. Ik heb generatieve AI gebruikt om mezelf weken werk te besparen bij codeerprojecten en ik vind het nuttig als een schrijfassistent om mijn gedachten te ordenen en lichtgewicht onderzoek uit te voeren - dat ik valideer voordat ik het gebruik, net als alles wat ik lees op internet.
De huidige modellen werken redelijk goed voor het samenvatten van inhoud in een artikel, het helpen schrijven en debuggen van applicatiecode, het maken van afbeeldingen en nog veel meer. Maar ze kunnen niet fungeren als een persoonlijke agent die me door een gemiddelde dag kan helpen. Daarvoor heeft de AI toegang nodig tot al mijn persoonlijke gegevens: e-mail, contactpersonen, agenda's, berichten, foto's en meer. ChatGPT kan niet "mijn vrouw mijn reisgegevens sturen voor mijn vlucht volgende maand" - en ik zou het ook niet de benodigde informatie geven.
Als Apple Intelligence op de markt komt, zullen de functies vooral bestaan uit een minder domme Siri, kleine aardigheidjes zoals e-mailoverzichten in berichtenlijsten, relatief saaie samenvattingen en salontrucs zoals Image Playground en Image Wand. Maar als alles goed gaat, kan Siri binnen een paar jaar gaan lijken op de persoonlijke digitale agenten uit sciencefiction, met name Apple's Knowledge Navigator. Ondanks de verbazingwekkende prestaties van Apple silicium, zullen sommige AI-gedreven taken altijd de cloud nodig hebben, wat de reden was voor Apple's werk aan het ontwerpen, bouwen en schalen van Private Cloud Compute. Apple wil dat we onze meest gevoelige gegevens toevertrouwen aan

zijn AI-platforms en erkent dat vertrouwen verdiend moet worden. De theorie is goed - wanneer de Apple Intelligence-functies beschikbaar worden, zullen we zien hoe de werkelijkheid zich verhoudt.

Vertaling: DeepL