

AI Safety Checklist

Qué revisar antes de pegar información en una herramienta de IA y antes de publicar lo que produjo.

Usar IA en marketing implica dos riesgos distintos: lo que le entregás y lo que publicás. Esta lista cubre ambos. Recorrela completa las primeras semanas; después se vuelve automática.

Antes de escribir el prompt: qué NO compartir

- Datos personales de clientes: nombres, correos, teléfonos, identificaciones, direcciones.
- Información financiera no pública: precios negociados, márgenes, presupuestos, proyecciones.
- Material bajo acuerdo de confidencialidad, propio o de un tercero.
- Credenciales, llaves de API, contraseñas o accesos de cualquier tipo.
- Documentos internos sensibles: estrategia sin publicar, temas legales, asuntos de personal.

Alternativas seguras

- Anonimizá: reemplazá nombres reales por «Cliente A», «Cliente B».
- Agregá: en vez de la base completa, pegá el resumen estadístico.
- Describí en lugar de pegar: contá la situación sin los datos identificables.
- Verificá la política de la herramienta sobre uso de tus datos para entrenamiento y desactivá esa opción si existe.

Durante el trabajo: verificación

- Toda cifra, fecha, nombre propio o cita se confirma en la fuente original.
- Toda URL que devuelva el modelo se abre antes de usarla.
- Ninguna afirmación legal, médica, financiera o de salud sale sin revisión de alguien calificado.
- Las comparaciones con competidores se sostienen con evidencia pública verificable.

Antes de publicar: revisión de marca y sesgo

- ¿El texto dice algo que solo tu marca podría decir, o serviría igual para un competidor?
- ¿Hay afirmaciones que la marca no puede sostener si alguien las cuestiona?
- ¿El material representa a las personas de forma justa, sin estereotipos de género, edad, origen o clase?
- ¿Se usaron imágenes generadas que puedan confundirse con fotografía real de personas o hechos?
- ¿Alguien con nombre y apellido se hace responsable de lo publicado?

Transparencia

No hay una regla universal sobre declarar el uso de IA, pero sí un criterio útil: declararlo cuando ocultarlo cambiaría la interpretación del público. Una imagen generada que parece documental exige aviso; un correo cuyo borrador redactó un modelo y una persona editó, no.

Regla de cierre

Nada se publica sin que una persona lo lea completo, lo entienda y esté dispuesta a defenderlo.

La herramienta produce; el criterio y la responsabilidad son humanos. Si no podés defender una frase, no es que falte editarla: es que no debería salir.