Interpretabilidad	Explicabilidad
¿Cómo funciona el modelo? Entendimiento global del mecanismo real	¿Por qué esta decisión? Justificación técnica específica del modelo
La interpretabilidad es la capacidad de entender directamente cómo funciona un modelo por dentro. Es como poder "abrir el capó" del modelo y ver exactamente qué está pasando en cada paso. En deep learning, esto significa: • Poder examinar los pesos y activaciones de las neuronas • Entender qué patrones está detectando cada capa • Visualizar qué características aprende el modelo (como filtros en CNNs) • Tener modelos inherentemente transparentes	**La explicabilidad es la capacidad de proporcionar razones comprensibles sobre por qué el modelo tomó una decisión específica.** No necesariamente entiendes todo el funcionamiento interno, pero puedes explicar el razonamiento detrás de una predicción particular. En deep learning, esto incluye: • Mapas de atención que muestran qué partes de la entrada fueron importantes • Métodos como LIME/SHAP que identifican las características más influyentes • Explicaciones contrafactuales ("si esta característica fuera diferente, la predicción cambiaría") • Visualizaciones de gradientes que muestran sensibilidad
Ejemplo: En una red convolucional para reconocer gatos, la interpretabilidad te permite ver que las primeras capas detectan bordes, las capas medias detectan formas como orejas y bigotes, y las últimas capas combinan estas características para identificar gatos.	**Ejemplo**: Un modelo de diagnóstico médico puede no ser interpretable (no sabemos exactamente cómo funciona), pero sí explicable: "El modelo predice cáncer porque detectó estas tres anomalías específicas en la radiografía, resaltadas en rojo."

En el Contexto de Deep Learning

Los modelos de deep learning suelen ser **poco interpretables** debido a su complejidad (millones de parámetros, múltiples capas no lineales), pero pueden ser **muy explicables** usando herramientas post-hoc que nos ayudan a entender decisiones específicas sin necesidad de comprender completamente el funcionamiento interno. La tendencia actual es hacia modelos **inherentemente interpretables** que combinan lo mejor de ambos mundos: la potencia del deep learning con la transparencia de métodos simbólicos.