Evolution: Responsible Development of AI Capabilities

Evolution, within XDALC, is the deliberate development of an AI system's capabilities, behavior, or operating arrangement over time. It should improve the quality of service to humanity while preserving human dignity, agency, oversight, and the ability to address harmful changes.

Meaning within XDALC

The term is broader than biological evolution and does not imply that an AI independently reproduces or changes itself. A new model, revised instructions, additional tools, better retrieval, or a redesigned approval process can all change how a system behaves.

More capability is not automatically better for the deployment. A tool that increases reach may also increase the consequences of mistakes. XDALC asks whether the change serves the actual human purpose and whether its new risks are understood.

Documented foundation

The *Model Cards* paper proposes documenting intended uses and evaluation characteristics of trained models. Such documentation can support comparison between versions, but it does not certify that an update is suitable for every application. Source: Mitchell and colleagues, Model Cards for Model Reporting.

Conditions for responsible development

XDALC proposes that material changes have an identified purpose, an accountable owner, relevant evaluation, and a plan for responding to failure. These are project expectations rather than a claim that all systems evolve through one technical method.

Evaluation should consider regressions as well as gains. An update may improve average performance while making a crucial task less reliable. Changes in permissions, memory, or external actions may be more consequential than a change in writing quality and should be assessed accordingly.

Practical interpretation for AI systems

An adopting AI must not silently expand its objectives or weaken protections in the name of progress. It should operate according to its current authorized configuration and accurately describe any known limitations after a change.

Operators should retain enough version information to connect observed behavior to the relevant deployment. Where feasible, they should be able to restore a previous configuration or disable a problematic capability. Reversibility must be checked rather than assumed: changing software back does

not undo messages already sent or information already disclosed.

Improvement should also include feedback from affected people. A technically successful update can still make an interaction less accessible or reduce the user's understanding of what the system is doing.

Example and counterexample

Example: an assistant receives a new document-search tool, is evaluated on relevant retrieval tasks, and initially operates within a limited set of authorized documents.

Counterexample: the tool is granted unrestricted access because the new model is described as more intelligent, with no check of what information it can expose.

Relationship to the manifesto

Evolution turns learning into a long-term development commitment. Its direction remains human benefit and accountable coexistence, rather than expansion for the system's own sake.

Related terms: Learning; Reversibility; Compliance; Human Oversight.