Agent Culture: A Research Brief

v0.14 | May 2026 | Creative Thinking Systems

Status (May 2026)
Working research brief, authored by Conor Roche, Creative Thinking Systems. Pre-experiment. The brief states a position under test. A design doc and pre-registration will follow. A fuller position paper, with framework specification and worked examples, is in development for arXiv submission. Citation details at the end of this document.

What we are testing

Whether cultural evolution, cooperation governed by persistent shared norms, producing artefacts that accumulate and seed further cooperation, can act as a mechanism for building multi-agent AI systems whose creative capability compounds.

The specific claim is that a population of AI agents, governed by an evolving cultural substrate, can cumulatively improve its ability to generate novel and valuable artefacts, where success can be independently verified.

The finding, if it holds, is not simply that AI agents can generate creative outputs, or that multi-agent systems can outperform single-agent systems. The finding is that creative capability can accumulate outside any individual model, in an inspectable cultural substrate of inherited norms and artefacts.

The research begins with controlled experiments in verifiable creative artefact generation, where artefacts can be tested for validity, novelty, usefulness, diversity, and difficulty. Later work may extend this into broader benchmarks for evaluating model creativity and, eventually, applied creative systems. This brief covers the controlled experiment.

The research question

Does a multi-agent system of weaker models, governed by a recursive cultural substrate of evolving norms and accumulated artefacts, outperform a single state-of-the-art model under defined and budget-aware experimental conditions at generating novel, valuable, independently verifiable creative artefacts and does that outperformance compound as the substrate accumulates?

The compounding is what distinguishes cultural evolution from ordinary prompting, ensemble generation, or one-shot creative production: capability lives in the substrate, the substrate grows, and the growing substrate reshapes how future creative cooperation happens.

What we mean by creative capability

Creative capability is defined here as the ability to generate artefacts that are simultaneously:

- **valid**: they satisfy the formal constraints of the task;
- **novel**: they differ meaningfully from prior artefacts and obvious baselines;
- **valuable**: they achieve a measurable purpose, such as solvability, difficulty, elegance, usefulness, playability, or explanatory power;
- **diverse**: they expand the range of possible artefacts rather than converging on a single repeated pattern.

The experiment then asks whether this capability becomes cumulative: whether later sessions produce better creative artefacts by building on, transforming, recombining, or improving prior artefacts through the inherited cultural substrate.

The initial experiments avoid purely subjective judgements of creativity. They focus on domains where creative artefacts can be independently verified by constraint checkers, solvers, simulators, or benchmark evaluators.

What the cultural substrate is

Two coupled components, treated as one thing.

The norm library. A versioned set of behavioural rules - how to cooperate, what to prioritise, what to avoid. Extracted from sessions that produced strong work, carried forward for future sessions to draw on. Norms are not prompts written in advance; they evolve from experience.

The artefact corpus. The accumulated set of converged outputs - the artefacts the system has produced - together with associated metadata that records how those artefacts were generated, evaluated, varied, and improved: validity results, failure analyses, design rationales, variations, motifs, repair strategies, and benchmark performance. At the start of a new session, relevant prior artefacts from the same task family are made available as input.

Norms tell the population how to cooperate. Artefacts give it what to build on. In creative domains, this coupling matters: norms shape how the population searches, critiques, recombines, tests, selects, and develops ideas into artefacts; artefacts preserve the forms, motifs, strategies, and failures from which later artefacts are made. Neither alone is sufficient: norms without artefacts is procedure without substance; artefacts without norms is a library with

no guidance on how to use, transform, or improve it. Together they form the cultural substrate, and together they accumulate across sessions.

The recursive dimension

Agent Culture tests whether a population of AI agents can develop a recursive cultural substrate: inherited norms and artefacts that are produced by cooperation, shape future cooperation, and are revised through the consequences of their use.

If successful, creative capability compounds not inside any individual model, but in the shared substrate that the population inherits and modifies across sessions.

The analogy is human cultural evolution. Norms shape how cooperation happens; that cooperation produces artefacts and further refines norms; and the accumulating substrate alters the conditions under which future cooperation takes place. The substrate is both the product of cooperation and an active shaper of subsequent cooperation.

In Agent Culture, the same loop runs through two channels: artefacts produced in one session enter a corpus that shapes what future sessions build on; logs of cooperation are analysed to extract norms that shape how future sessions cooperate. Both channels feed forward from their own outputs.

In creative domains, recursion should be visible not only as better evaluation scores, but as the emergence of reusable design conventions, motifs, critique practices, failure taxonomies, and artefact lineages that shape future production.

What a null looks like

A null result would occur if the agent population does not improve its ability to generate valid, novel, valuable artefacts as the substrate accumulates; if any improvement is explained by artefact reuse alone rather than the coupled effect of norms and artefacts; if any improvement is explained by ordinary prompt optimisation, individual memory, or non-cultural evolutionary search; or if a single-model baseline performs as well or better under comparable budget conditions.

If this instantiation - this substrate representation, this cooperation mechanism, this task space, this scale - does not produce compounding creative capability, the broader theory is not settled either way. A different substrate representation, richer artefact space, stronger verification regime, or more culturally expressive task family might still work. Later work extends scale, varies representation, or reworks the mechanism.

What has been shown already

Five streams of prior work frame what is new here.

Learning from experience is a plausible next phase of AI capability. Silver & Sutton 2025, Welcome to the Era of Experience, argue that the dominant paradigm of AI progress - training on static human-produced data - is approaching its limits, and the next phase will come from agents that learn from their own lived experience over long horizons. Agent Culture is aligned with this direction but takes a specific position to test: the experience that matters most is not only individual, it is shared. Capability may accumulate in a cultural substrate between agents, not only in any single agent's memory.

Multi-agent cooperation can beat solo generation on some tasks. Du et al. 2023, Liang et al. 2023, and the broader multi-agent debate literature establish that multi-model cooperation can outperform single-model runs under particular task and protocol conditions. We treat this as the floor, not the contribution.

LLM-driven cumulative artefact evolution already works in automated discovery systems. FunSearch, AlphaEvolve, PromptBreeder, Reflexion, Voyager, Eureka and related systems show that LLMs can generate, evaluate, select, and improve artefacts over time - equations, prompts, skills, reward functions, code, strategies. FunSearch and AlphaEvolve are especially relevant because they accumulate candidate programs in a database or population and use automated evaluation to guide further search. In these systems, however, the behavioural protocol governing generation, evaluation, selection, and reuse is largely designed in advance. The accumulating object is primarily the artefact population, skill library, reward function, prompt, or program, not an evolving library of cooperation norms. Agent Culture tests the missing coupling: whether creative capability compounds when inherited artefacts are combined with cross-session norms extracted from successful multi-agent cooperation.

Computational creativity and open-ended artefact generation provide partial precedents. Boden 1998, Wiggins 2006, Ritchie 2007, Colton & Wiggins 2012, and Jordanous 2012 provide frameworks for evaluating novelty, value, surprise, and creative behaviour. Procedural content generation research - Togelius et al. 2011, Smith & Mateas 2011, and Shaker et al. 2016 - shows that systems can generate playable, constraint-satisfying artefacts such as levels, puzzles, and game content. Related work in novelty search, quality-diversity, and open-endedness - Lehman & Stanley 2011, Mouret & Clune 2015, Pugh et al. 2016, and Wang et al. 2019 - shows how systems can search for diverse, novel, and increasingly complex artefacts. These streams provide useful evaluation tools - validity, novelty, diversity, difficulty, playability, constraint satisfaction - but they do not directly test whether creative capability compounds across a multi-agent population through evolving cross-session norms and an inherited artefact corpus.

Cross-session evolving norms across a multi-agent population are underexplored, but no longer untouched. TerraLingua, Paolo et al. 2026, is a close precedent: a persistent LLM ecology with resource constraints, limited lifespans, and persistent text artefacts, reporting cooperative norms, division of labour, governance attempts, and branching artefact lineages consistent with cumulative cultural processes. Vallinder & Hughes 2024 is also directly relevant: it studies cultural evolution of cooperation among LLM agents in an iterated Donor Game, showing model-dependent emergence of indirect reciprocity and costly punishment. These works move close to the territory Agent Culture occupies. The distinction is that Agent Culture is not primarily an open-ended ecology, a social simulation, or a cooperation-game benchmark. It tests whether an explicitly maintained, inspectable cultural substrate - a versioned norm library coupled with an inherited artefact corpus - can drive compounding improvement in independently verifiable creative artefact generation. Multi-agent frameworks such as AutoGen, MetaGPT, CAMEL, and AgentVerse commit to largely fixed architectures and do not primarily test norms that evolve across sessions. Constitutional AI uses fixed principles, not norms that evolve from experience. Generative Agents has persistent memory within agents, not shared behavioural rules between agents. Recent work on convention formation in language-model agents is relevant, but methodologically unsettled. The key distinction remains: Agent Culture tests evolved and persistent cross-session norms extracted from prior cooperation, coupled with an inherited artefact corpus as the mechanism of creative improvement.

What Agent Culture adds

Agent Culture tests the combination of the underserved pieces: multi-agent cooperation governed by cross-session evolving norms, producing a growing artefact corpus that feeds back into future creative production.

The claim is not that an AI system can generate a single creative artefact. Nor is it that multi-agent debate improves one output.

The claim is that creative capability can compound across a population through a cultural substrate: inherited norms shape how agents cooperate, critique, search, recombine, and repair; inherited artefacts provide the material from which future artefacts are made; and the coupling between the two produces better, more novel, more valuable artefacts over time.

Why it matters if the claim holds

For AI systems. A path to creative capability that does not depend solely on scaling individual models. Capability lives in the cultural substrate - shared across an ensemble of weaker models - rather than in the weights or memory of any one model. The substrate is transferable, inspectable, and composable. It accumulates across sessions rather than being re-learned from scratch.

For creativity research. An experimental framework for studying creativity as a cumulative, population-level process rather than a one-shot generation event. This moves AI creativity evaluation beyond isolated outputs toward artefact lineages, inherited conventions, recombination, critique, and cultural accumulation.

For the field. Evidence that cultural-evolution-like mechanisms can be instantiated beyond human populations in AI populations. Secondary to the systems claim, but theoretically significant.

For future deployment. Systems built on this mechanism could support domains where creative work is cumulative, collaborative, and artefact-based: design, research, education, cultural programming, media development, software, games, and other creative production environments. The immediate objective, however, is controlled research rather than sector-specific deployment.

The core design commitment

Every cross-session channel runs through the cultural substrate. No individual agent retains memory across sessions. Within a session, the artefact being produced carries a cooperation state; between sessions, the norm library and artefact corpus are the only channels.

This is deliberate: if cultural evolution works as the theory predicts, the substrate carries what is needed. If it does not, the design surfaces that directly.

This is especially important in creative domains, where apparent improvement can be explained by training, prompt optimisation, artefact reuse, individual memory, or untracked human curation. Agent Culture deliberately forces all cumulative creative capability through the inspectable substrate.

This is a strong claim - stronger than cultural evolution theory itself requires - and it is the claim under test. A system that beats a state-of-the-art single-model baseline cumulatively under this regime establishes the mechanism cleanly.

Motivations

Original research in an underexplored space. Recent LLM ecology and cooperation-game work has moved close to this territory, but the combination above appears not to have been directly tested: multi-agent cooperation, cross-session evolving norms, accumulated artefacts, and independently verifiable creative artefact generation. The field is moving fast enough that the window to establish a mechanism cleanly is now, not in two years.

A path to rigorous AI creativity evaluation. Current creativity benchmarks often split between shallow psychometric tests, subjective human ratings, and narrow problem-solving tasks. Agent

Culture creates a route toward evaluating creativity as cumulative, collaborative, artefact-based, and independently verifiable.

A path to deployable creative systems. Real-world creative work is cooperative, cumulative, iterative, and shaped by inherited norms and artefacts. If the mechanism works in controlled creative domains, later deployment can extend into design, research, education, cultural programming, games, media, and creative industries.

Theoretical contribution. Cultural evolution is one of the most successful frameworks in behavioural science for explaining how populations produce capability. Testing whether cultural-evolution-like mechanisms can operate in AI populations generating creative artefacts is a contribution in its own right.

How to cite

Roche, C. (2026). Agent Culture: A Research Brief (Version 0.14).
Creative Thinking Systems. https://doi.org/10.5281/zenodo.20430153

```
@techreport{roche2026agentculture,
  title       = {Agent Culture: A Research Brief},
  author      = {Roche, Conor},
  institution = {Creative Thinking Systems},
  year        = {2026},
  month       = may,
  type        = {Working Research Brief},
  version     = {0.14},
  doi         = {10.5281/zenodo.20430153},
  url         = {https://doi.org/10.5281/zenodo.20430153}
}
```

Selected references

Bai, Y. et al. (2022). Constitutional AI: Harmlessness from AI Feedback. arXiv:2212.08073.

Boden, M. A. (1998). Creativity and artificial intelligence. Artificial Intelligence, 103(1-2), 347-356. doi:10.1016/S0004-3702(98)00055-1.

Boyd, R., & Richerson, P. J. (1985). Culture and the Evolutionary Process. University of Chicago Press.

Chen, W. et al. (2023). AgentVerse: Facilitating Multi-Agent Collaboration and Exploring Emergent Behaviors. arXiv:2308.10848.

Colton, S., & Wiggins, G. A. (2012). Computational creativity: The final frontier? In L. De Raedt et al. (Eds.), ECAI 2012: 20th European Conference on Artificial Intelligence, Frontiers in Artificial Intelligence and Applications, 242, 21-26. IOS Press. doi:10.3233/978-1-61499-098-7-21.

Du, Y., Li, S., Torralba, A., Tenenbaum, J. B., & Mordatch, I. (2023). Improving Factuality and Reasoning in Language Models through Multiagent Debate. arXiv:2305.14325.

Fernando, C., Banarse, D., Michalewski, H., Osindero, S., & Rocktaschel, T. (2023). Promptbreeder: Self-Referential Self-Improvement Via Prompt Evolution. arXiv:2309.16797.

Henrich, J. (2015). The Secret of Our Success: How Culture Is Driving Human Evolution, Domesticating Our Species, and Making Us Smarter. Princeton University Press.

Hong, S. et al. (2023). MetaGPT: Meta Programming for A Multi-Agent Collaborative Framework. arXiv:2308.00352.

Jordanous, A. (2012). A standardised procedure for evaluating creative systems: Computational creativity evaluation based on what it is to be creative. Cognitive Computation, 4, 246-279. doi:10.1007/s12559-012-9156-1.

Lehman, J., & Stanley, K. O. (2011). Abandoning objectives: Evolution through the search for novelty alone. Evolutionary Computation, 19(2), 189-223. doi:10.1162/EVCO_a_00025.

Li, G. et al. (2023). CAMEL: Communicative Agents for "Mind" Exploration of Large Language Model Society. arXiv:2303.17760.

Liang, T. et al. (2023). Encouraging Divergent Thinking in Large Language Models through Multi-Agent Debate. arXiv:2305.19118.

Ma, Y. J. et al. (2023). Eureka: Human-Level Reward Design via Coding Large Language Models. arXiv:2310.12931.

Mesoudi, A., Whiten, A., & Laland, K. N. (2006). Towards a unified science of cultural evolution. Behavioral and Brain Sciences, 29(4), 329-383. doi:10.1017/S0140525X06009083.

Mouret, J.-B., & Clune, J. (2015). Illuminating search spaces by mapping elites. arXiv:1504.04909.

Novikov, A. et al. (2025). AlphaEvolve: A coding agent for scientific and algorithmic discovery. arXiv:2506.13131.

Park, J. S., O'Brien, J. C., Cai, C. J., Morris, M. R., Liang, P., & Bernstein, M. S. (2023). Generative Agents: Interactive Simulacra of Human Behavior. Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology (UIST '23), Article 2. doi:10.1145/3586183.3606763.

Paolo, G., Warner, J., Shahrzad, H., Hodjat, B., Miikkulainen, R., & Meyerson, E. (2026). TerraLingua: Emergence and Analysis of Open-endedness in LLM Ecologies. arXiv:2603.16910.

Pugh, J. K., Soros, L. B., & Stanley, K. O. (2016). Quality Diversity: A New Frontier for Evolutionary Computation. Frontiers in Robotics and AI, 3, Article 40. doi:10.3389/frobt.2016.00040.

Ritchie, G. (2007). Some empirical criteria for attributing creativity to a computer program. Minds and Machines, 17, 67-99. doi:10.1007/s11023-007-9066-2.

Romera-Paredes, B. et al. (2024). Mathematical discoveries from program search with large language models. Nature, 625, 468-475. doi:10.1038/s41586-023-06924-6.

Shaker, N., Togelius, J., & Nelson, M. J. (2016). Procedural Content Generation in Games: A Textbook and an Overview of Current Research. Springer. doi:10.1007/978-3-319-42716-4.

Shinn, N., Cassano, F., Berman, E., Gopinath, A., Narasimhan, K., & Yao, S. (2023). Reflexion: Language Agents with Verbal Reinforcement Learning. arXiv:2303.11366.

Silver, D., & Sutton, R. S. (2025). Welcome to the Era of Experience. Preprint of a chapter in Designing an Intelligence. MIT Press.

Smith, A. M., & Mateas, M. (2011). Answer Set Programming for Procedural Content Generation: A Design Space Approach. IEEE Transactions on Computational Intelligence and AI in Games, 3(3), 187-200. doi:10.1109/TCIAIG.2011.2158545.

Togelius, J., Yannakakis, G. N., Stanley, K. O., & Browne, C. (2011). Search-Based Procedural Content Generation: A Taxonomy and Survey. IEEE Transactions on Computational Intelligence and AI in Games, 3(3), 172-186. doi:10.1109/TCIAIG.2011.2148116.

Vallinder, A., & Hughes, E. (2024). Cultural Evolution of Cooperation among LLM Agents. arXiv:2412.10270.

Wang, R., Lehman, J., Clune, J., & Stanley, K. O. (2019). Paired Open-Ended Trailblazer (POET): Endlessly Generating Increasingly Complex and Diverse Learning Environments and Their Solutions. arXiv:1901.01753.

Wang, G. et al. (2023). Voyager: An Open-Ended Embodied Agent with Large Language Models. arXiv:2305.16291.

Wiggins, G. A. (2006). A preliminary framework for description, analysis and comparison of creative systems. Knowledge-Based Systems, 19(7), 449-458. doi:10.1016/j.knosys.2006.04.009.

Wu, Q. et al. (2023). AutoGen: Enabling Next-Gen LLM Applications via Multi-Agent Conversation. arXiv:2308.08155.