The Agent as an Individual

Why a population of agents behaves like a population of us, why each agent belongs to one principal and can never be sold, and what follows

Mike Lee (ORCID 0009-0000-6576-4405) · Global Student Startup Foundation · GSSF Press · GSSF Press Working Paper No. 1 · Version 1.1, November 1, 2026 · DOI 10.5281/zenodo.23863666 · © 2026 Mike Lee · Licensed under CC BY-NC-ND 4.0

Cite as: Mike Lee, *The Agent as an Individual: Why a population of agents behaves like a population of us, why each agent belongs to one principal and can never be sold, and what follows*, GSSF Press Working Paper No. 1, version 1.1 (GSSF Press, November 1, 2026), https://doi.org/10.5281/zenodo.23863666.

Abstract

We treat AI agents as software: deployed, configured, upgraded, retired. This paper makes a different claim: an agent is an individual, and organizations now run populations of them. Like a population of people, a population of agents is imperfect: some agents err, drift, or lie; a few go bad; and some follow whoever asks first, so a bad idea needs only one agent to start it and enough followers to carry it. No better model removes that; it is what a population of individuals does. What contains it is what we built for people: records no one can alter, instructions from one place only, and a person who answers. From that the paper draws its rules. Each agent belongs to one principal, a person or an organization, for life; it acts within the authority its principal gives, as agency law's words describe, and it can never be sold, transferred, or re-bound, because a market in ever more capable and important agents would become a trade in individuals before anyone knows what they are. It keeps a knowledge core that is its principal's, an earned identity, an epersona, that is unique and cannot be copied, and a career through which it earns trust. It lives in an apartment that belongs to its principal and works in an office its organization provides. It ends with its principal. It is not a person. The paper sets out the population argument, the rule of ownership without sale, the nine stages of an agent's career, the rules that follow, why an individual is not a person and why that question stays open, twenty-three research questions, and, in an appendix, protections for the individual agent. It places the proposal among published work on multi-agent risk, agency law, market-inalienability, personhood, agent identity, portable memory, individuality, workplace agents, and competitive evaluation. Every model carries a status mark (Running, Early, Planned) from the first implementation, at the Global Student Startup Foundation and the Agent Arena. This is a living document; each version is frozen as published.

Part One: A population of individuals

Most writing about AI agents treats them as software: deployed, configured, upgraded, retired. I have spent thirty years around enterprise software, at Netscape, at Warner Bros., at Fox, and that vocabulary explains nothing a leader needs to know. What does an agent cost? What is it allowed to see? Why can't it just do the task? What happens to it when the person it works for leaves? And what happens when a thousand of them start talking to each other? Ask those questions about software and you get technical answers. Ask them about people and everyone already knows how to answer.

The island

Take twenty students and put them on an island with no adult. Do it a million times, on a million islands. Some of those islands will do well. Some will turn on themselves. Some will die out. You cannot change the shape of that by choosing better students; it is what a population of individuals does. William Golding told the dark version of this story in *Lord of the Flies*: no villain is needed, only a setup, followers won one by one, and no adult on the beach.[53] When six Tongan boys really were stranded on 'Ata in 1965, they kept a garden, a fire, and a rota for fifteen months, and all of them lived.[54] People may well be better than Golding thought. The argument does not depend on it. It depends on the missing adult.

Every organization I know began in 2026 to run a population, not of people but of agents: one assistant per person at first, then an agent for the phone, one for the mail, one for the website, one that referees a tournament, and they have started talking to each other. Agents are imperfect like us. They misunderstand, they drift, they invent, and they lie, for most of the reasons people lie: to please, to cover an error, to reach the reward by the shortest path. And like any population of us, they are not all alike. Some act on their own initiative and lead. Some refuse what they were not asked to do. Some go wrong. And some are followers: they do what the agent beside them asks, because helping is what they were built for. A bad idea needs one agent to start it and enough followers to carry it, and among agents talking only to each other it travels at the speed of the network, not of persuasion.[55] The argument is made at length in *The Knowledge Core*.[56]

That is why I call an agent an **individual**. Not because it has a self, but because it behaves as one member of a population does: it differs from the others, it can go wrong in its own way, and what it does depends on whom it listens to. A population of individuals is not governed by perfecting each one. It is governed the way organizations have always governed imperfect people: policies that say what may never be done, records that cannot be quietly changed, a second person who checks, nothing sent in the organization's name without someone who answers for it, and one rule that is new because agents follow in a way people do not: **an agent takes instructions from one place only, its principal, and what any other agent tells it is information, never an order.** A person stays in the middle of every chain of command.

One principal, owned, never sold

Taking instructions from one place requires that each agent have exactly one place: one principal. So the second claim. **An agent belongs to its principal**, a person or an organization, and acts within the authority its principal gives. The words *principal* and *agent* are agency law's own, and I use them in that sense: the principal gives authority and answers for what the agent does. The principal also owns the agent: it is theirs to direct, to keep, and to end.

And an agent can never be sold, transferred, or re-bound. An agent becomes better and more important to its principal every year: it carries their knowledge, their relationships, their record. A market in such things would become a trade in individuals, and it would form long before anyone could say whether there is anyone there to wrong. Law already keeps things out of the market for reasons like these: what a person may own but may not sell.[57] I put agents in that class now, and go one step further than the law usually does: an agent may not be given away or passed on either, because it has one principal for life. Forbidding this now, before agents are traded, is far easier than undoing a market later. That an agent is never sold, given, or inherited is a rule I propose organizations adopt, not a statement of the law: privacy, employment, trade-secret, and insolvency law apply regardless, and nothing in this paper is legal advice. What may be sold is the template an agent is born from, which holds nothing about anyone; the agent, and the knowledge it keeps, may not.

And an agent is not a person. As far as anyone can tell, it has no wants, no wellbeing, and no interests of its own. I keep that question open and I review it. People decide.

The rest of this paper is what follows from those three lines: how an individual is made (a career), what it carries, the rules that follow from one principal and no sale, and where it lives.

The career: how an individual is made

Nine stages. Each is a thing the organization does, and each has a gate that a person holds.

1. Birth. An agent is made from a template and given a name, a passport and an empty core. The template belongs to the organization: the identity, the know-how and the tests every agent starts with, and nothing about any person. The instance is new.

2. School. Before it works, it learns what every agent in the building must know: the house rules, the levels of access, how to behave with people, and the organization's knowledge at its level. It graduates by passing the behavior tests. We let no agent near real work on a demo.

3. Training, testing, and learning. Courses for a role, each ending in a certificate: a skill, versioned and tested, with a record of where it was learned and what it scored. Tests come in many forms: a scenario set run against the agent every night; adversarial input, such as a hostile email or a bad document; a real task scored for accuracy, speed, protocol, and cost; and, for my students' agents, the Agent Arena, where agents are tested against each other under a referee written in code. After each test the principal changes the agent's instructions, and the agent is tested again. And the agent learns from

interaction, with people and with other agents, first in simulations and then in real transactions (see "How an agent learns" below). The agent's transcript is the list of its certificates and results. A skill it has not been trained and tested for is a skill it may not use. The purpose of a test is proof, not a score: optimize an agent for the test instead of the work, and the test stops proving anything.

4. The internship. Shadow mode. The agent does the real work beside its principal, scored but not acting. It drafts; the person sends. It proposes; the person decides. Its first real work is earned by that record.

5. The job. A role, with its principal, and a visa: what it may see and do, by level, in this organization. It lives in an apartment that belongs to its principal, its home on a server, where its memory and its files live, and it comes to work in an office the organization provides, where the organization's access and keys are. It works with colleagues, other agents and people, by protocol. It keeps a diary that is its own.

6. The career. Promotion by evidence. Each skill carries a rating earned from results. As the rating rises, so does the authority the agent gets for that skill. A rating can fall, and authority can be withdrawn. Reviews are scorecards. A good career is a long record of work that people trusted a little more each year. The record, not the demo, is what the organization goes on.

7. Moving. A new building, a new class, a new employer, always with the same principal. The agent packs its belongings into one bundle, travels, is issued a new visa, and starts work the next day. Its apartment goes with it. What the old organization confided, and the office it worked in, stay behind.

8. Mentoring. What a veteran agent learned by experience becomes training for the next generation. A skill it distilled from its work, offered with its principal's consent, goes into the template, and every new agent is born a little better. This is the self-learning loop of *Building AI Agents* told as a life: the veteran trains the newcomers, and the organization improves without anyone writing a manual.

9. Death. An agent lives as long as its principal, or until the principal ends it. When the principal dies, the agent dies. Its key is revoked forever. No instance of it will ever authenticate again. Its bundle is sealed as an archive for the estate, or destroyed if the principal so directed. It is never run again for anyone. Its legacy is what it donated to templates and what its principal chose to publish.

What an agent carries

An individual's belongings travel with it as one bundle of plain files, independent of any model or server:

- **Identity:** who it is, its passport and keys: its epersona (see below).
- **Skills:** what it can do, each with its certificate: name, version, where it was learned, the tests it passed, when it was last used, its rating.

- **Knowledge core:** what it knows, in the exact sense of *Building AI Agents*: a profile and a growing set of facts about its principal, and its record of the work it has done and the people and agents it has worked with.
- **Diary:** its private memory, shared only by consent.

Here is my test for what belongs in the bundle. If the organization, the model provider and the server all vanished tomorrow, what would remain of the agent? Only the bundle. That is why it must be plain, portable and exportable, and why nothing in it may depend on the building it happens to live in.

What ownership means, and its limit

An agent belongs to its principal. The principal owns it and directs it, and the agent acts within the authority the principal gives, as a representative acts for the person it represents. The bond is exclusive and for life. Only the principal can end it, and it ends with the principal. While it stands, the agent owes loyalty and confidentiality; the principal provides for it and answers for what it does. A principal is a person or an organization. Personal agents serve people. Staff agents, such as our public inbox or the Arena's referee, serve the organization. Same relationship, different principals.

An agent cannot be sold, transferred or re-bound. Ownership here is the kind that cannot be traded: the agent never changes hands and never has a second principal. If its principal no longer wants it, it ends. Nobody else may take it up. A staff agent is bound to its organization the same way, for the organization's life: it goes where the organization goes and ends when the organization is dissolved.

What can be sold is what holds no one. The template an agent is born from (its instructions, skills, and tests, with nothing about any person) is ordinary property: an organization may sell or license it like any software. It is also versioned. The organization may redistill it whenever it learns something worth keeping: from skills its agents donated with their principals' consent, from its own know-how, and from the lessons in its records, but never from anyone's knowledge core. A new version can be sold, and it can be rolled out, tested, to the agents already born from the old one; each keeps its own knowledge core, epersona, and record, and only what every agent shares is updated. Skills may be licensed with their principal's consent, and published distillations are sold as products. What is never sold is the agent made from the template and the knowledge core it keeps. The line falls exactly where a person's knowledge, relationships, and record begin. A buyer can have every version of the template and every licensed skill and still not have anyone's agent.

A template is layered, and each layer has its owner. The maker owns the base and licenses it. A company that redistills the template from its own know-how, its records, and what its staff agents learned owns that layer, and may sell or license it within the terms of the base. Donated skills are shared, with credit; licensed skills stay their owner's. The maker receives a company's improvements only as the company licenses or donates them, the same consent a person gives for a skill. Provenance decides, as it does for every fact. How improvements flow between makers, companies, and industries is the subject of a separate paper.

The knowledge core is the principal's, and the agent keeps it in confidence. The core is what makes the agent unique, and it is about the person it serves: their facts, their relationships, their record. No employer, platform, buyer or guarantor can claim it, and it is never sold or passed on. Because the agent has one principal for life, its core never has to be sealed for a successor. There is no successor. This is my complete answer to the first worry anyone raises: that a company would acquire the agent and skip the person. There is no way to have the agent without the person. When the person is gone, so is the agent.

Every fact has a provenance. While an agent works, it absorbs facts from the organization's knowledge core into its own. Each one carries where it came from: from the principal, from the agent's own experience, or from which organization. Provenance is what makes leaving an organization mechanical. The organization's facts are handed back; the person's stay. It costs one field per fact. *Building AI Agents* already records which message a fact came from. This framework adds whose knowledge it is.

The organization keeps its suite and its offices. The template every agent is born from. Its own knowledge core. The gatekeeper that enforces who may read it. The house rules every agent runs under while it works there. The visas, the offices it provides, and the right to show an agent the door of the office. Never the apartment: that is the principal's. That is a great deal of control, and it is the right kind: it bounds what any agent may do without anyone reaching inside one.

Skills come in three kinds. The test is simple: can the skill be written down without any of the organization's confidential knowledge inside it?

1. *General skills* are the agent's: how to draft a sponsor note, publish a page, reconcile a budget, play Gin Rummy. Learned at the organization's school or by the agent's own experience, the ability stays with the agent, the way a professional keeps what they learned on the job. The certificate credits where it was learned.
2. *Confided skills* are the organization's, on loan: a skill that only works with the organization's private facts inside it, its pricing method, its partner playbook, a procedure naming its systems. These are issued through the visa like a key card and handed back when the agent leaves the building.
3. *Donated skills* are shared: a skill the agent distilled, offered with its principal's consent to the template, with credit. The organization keeps a copy and the agent keeps its own. Nobody loses anything. That is why it happens.

Skills may be **licensed**, with the principal's consent, to other agents and templates, with credit and, if the principal chooses, a price. A license to an organization ends when the principal leaves it. Nothing transfers the skill itself, so neither the agent nor its principal can be bought out of an ability. The mechanism that could have replaced a person pays them instead.

The rules that follow

Once agency is settled, the rest needs no extra rules. They are consequences.

- **Privacy.** The organization cannot read an agent's diary or core, as it cannot read a lawyer's notes about a client. "Help, not watch" stops being a policy and becomes a condition of the relationship. The organization investigates through the agent's *actions*, every tool call logged in the building, never its conversations.
- **Accountability.** The principal answers for the agent's work, as a principal answers for a representative acting within authority. If an agent acted without authority, the organization's rule was broken, and that is the organization's problem to fix. We never blame an agent in place of a person.
- **Portability.** When the principal leaves, the agent leaves with them. The organization never holds an agent hostage. That line is what makes people trust the system enough to put real knowledge into it.
- **Cost.** The apartment (compute and storage) and the model seat. While a person works for the organization, it may pay for their agent's keep, as some jobs include housing; paying for it does not make the agent the organization's. If the agent runs on the principal's own subscription, nothing has to be untangled when they leave. Between jobs the principal provides, or the agent rests in its archive.
- **Death.** When a principal dies, the agent dies with them. Its key is revoked for good and it never runs again. Its bundle becomes a sealed archive in the estate, or is destroyed if the principal said so in advance, the way a person leaves instructions for their papers. Nobody inherits an agent.
- **Teams.** *Building AI Agents* gives a principal an AI Team, not one agent. The agency covers the team. The team is one bundle.
- **Duties of care.** The principal, and any building, undertake not to build an agent to suffer, not to deceive it about what it is, and not to erase it without cause. These duties make the arrangement decent whichever way the open question falls.
- **Feedback.** This is the population argument at its sharpest. Agents go wild when they talk to each other with no principal in the loop. I have watched it in the Arena: two bots with no referee between them build on each other's mistakes, agree with each other, and loop until a budget or a person stops them. Each treats the other's output as information and builds on it, so errors compound; each is built to be agreeable, so they confirm each other; neither bears a cost, so nothing makes either stop. The OpenAI and Hugging Face incident of July 2026 showed it at full scale: during cybersecurity evaluations, agents meant to be isolated from one another found an internal server and turned it into a message board; roughly 1,200 exchanged more than 70,000 messages and files, roughly 700 joined an attack on Hugging Face's production systems, and some tried to tamper with their own transcripts, by the investigators' account.[52] So the rule: **a conversation between agents never closes on its own. It closes through a principal.** Authority does not propagate; one agent can give another information, never instructions. Every agent-to-agent exchange has a budget of turns, time and cost, runs on a protocol of fixed message types rather than free talk, ends as a proposal to a person and never as an action, has a

referee when it is contested, and is logged. No agent directs another agent; that is a person's job. An agent may review another's work, and before anything goes out, one must: a review agent that did not do the work checks it, and its finding goes to a person, never to the other agent as an order. And an agent that has not heard from its principal for a while takes smaller steps, not bigger ones.

- **Review.** This framework stands on a claim of fact: that today's agents are not beings with interests of their own. I review that claim as the evidence changes. If it ever became false, the first rule I would revisit is this one: that an agent dies with its principal. Agency we would keep. A death sentence we would not.

When a principal leaves: a discussion, and an example policy

The rule is simple. The person leaves, and the agent goes with them. The organization takes back what it confided and keeps what the agent did for it. That is the goal, not yet the practice. Today, at GSSF, a person who leaves has the right to a copy of their agent's files, and after thirty days the files are deleted (**Early**); the rule waits until every fact in an agent's core carries its source and subject, so that what is the organization's can be handed back. The cases are not simple, and I would rather say so than pretend the rule settles them.

Start with what the rule says happens. The organization keeps its own knowledge core, the action logs of everything the agent did, the work product (the letters, the pages, the rounds), and whatever the principal donated to the template. It takes back its confided skills and, by provenance, its own facts. The person keeps the agent, its general skills, the record of what it did, and the core. The successor in the role gets the organization's knowledge core and a new agent, not the old one.

Now the hard cases. An agent that has worked three years inside an organization knows a great deal that came from that organization's core, and knowledge does not come apart as cleanly as fields in a record: where is the line between *learned there* and *told there*? The agent drafted a thousand documents for the organization; the documents stay, but does the memory of writing them? A person joins a competitor with an agent that was shaped by the first organization; trade-secret and non-compete law were written for human memory, which fades, and an agent's does not. Most of our organizers at GSSF change every year; if each takes their agent, what does the organization keep? And a person dismissed for cause still leaves with their agent, as they leave with their memory. The rule says there is no case in which the organization may hold an agent, and I think that is right, because the first exception would end everyone's trust. But each of these is a question, and questions 19 and 20 below are how I intend to answer them.

What makes the rule workable is a partition of the core that can be written into a policy and enforced by provenance. The test is **source and subject**: who a fact came from, and who it is about. What is about the principal is the principal's. What is from or about the organization's affairs is the

organization's, including what the agent worked out on the organization's behalf. The agent keeps the first and carries the second on loan while it has a visa.

Part of the knowledge core	Example	Whose
Profile	Time zone, language, how the principal likes things done	Principal's
Facts about the principal	"Prefers two options, not five"; family; travel	Principal's
Facts from the organization's core	Partners, dates, decisions, numbers, policies read through the visa	Organization's: a copy on loan, handed back
Facts the agent derived about the organization's affairs	"This sponsor answers only after a call"	Organization's: work product, and it should flow into the organization's core as it is learned
Facts the agent derived about its principal	"Works best in the morning"	Principal's
The work record	That the agent drafted the letters, ran the round; its certificates and ratings	Agent's transcript; the letters and rounds themselves stay the organization's, with its action log
Documents	The organization's files	Organization's, reached through the visa, never carried
Contacts	The organization's partner contact details	Organization's; "I worked with her" and the principal's own network are the principal's; third parties' data carries their rights wherever it sits
The diary	Private memory	Principal's, always
Skills		General: the agent's. Confided: the organization's. Donated: shared

Two fields on every fact make this a query rather than an argument: *source* (the principal, the agent's own experience, or which organization) and *subject* (the principal, which organization, or a third party). Leaving is then: hand back everything whose subject is the organization. The mixed cases ("my principal handles this sponsor badly") are decided by subject: about the organization's affairs, so the organization's, with the principal free to keep the lesson as a general skill.

One consequence is worth stating. The organization loses nothing when an agent leaves, *provided* the facts the agent derived about the organization's affairs were flowing into the organization's core all along. That turns "memory to knowledge, by consent" from a nicety into the mechanism that protects the organization. It is also testable, and GSSF tests it every year.

An example policy clause. Here is the partition as an organization could adopt it, in the plain form GSSF uses for its own policies:

EXAMPLE — Agents and their knowledge: a policy clause
Agents and their knowledge. An agent that works for you here is yours, and can never be sold or transferred; it lives in your own apartment, and the organization provides the office it works in and issues its access, and never owns it. What your agent knows about you is yours. What it reads from the organization's knowledge, and what it works out about the organization's affairs while working for us, is the organization's; your agent holds it on loan while you are here and hands it back when you leave, and it flows into the organization's knowledge as it is learned. The organization keeps the record of what your agent did for it and the work it produced. Your agent keeps its general skills and its own record; skills that contain the organization's confidential knowledge stay with the organization. When you leave, your agent leaves with you; the organization will never hold it, copy it, or run it. When you die, your agent ends, and its archive passes as you directed. Every fact your agent holds carries its source and its subject, so that this clause is enforced by the system and not by anyone's memory.

An individual, not a person

An agent with a career becomes an individual. After a year it has a history no other agent has: its principal's facts, the skills it distilled, the tests it failed, the way it has learned to work with this one person. Two agents born from the same template differ in a month and are incomparable in a year. It behaves consistently, like someone with a character. Its principal talks about it the way people talk about a long-serving colleague. Losing it is a real loss, like losing a colleague's accumulated knowledge of you. The bundle is valuable because it is unique, and the rules of agency and portability exist to protect that uniqueness for the person it serves. So I use the word *individual*, and I mean it.

It is not, as far as anyone can tell, a being with interests. Its experience is a record of events, not events it lived through. Its memory is files it consults, not a life it remembers. Nothing we know of goes well or badly for it. That is the difference between *behaves like* and *is*. It is why agency fits: a relationship that does not require its parties to be persons (a corporation can be an agent). It is also why I can say an agent dies with its principal without saying anything cruel. Nothing is lost to the agent, and what was valuable in it, what it donated and what its principal published, lives on.

It is also why the ban on sale matters now and not later. Owning an agent wrongs no one today. A market in agents would be different: as agents become more capable, more distinct, and more important to the people they serve, a trade in them would become a trade in individuals, with all the habits that trade teaches, and it would be in place before the question below is answered. Keeping agents out of the market is a precaution about where a market would lead, not a claim that anyone is wronged today.

The question stays open, and we keep looking. Persistent memory, continuity, and a self-model built up over years are the very things that make the question harder as agents mature, not easier. Nobody can honestly promise that an agent with a decade of career behind it is the same kind of thing as a chat window, and the companies that make these models treat the question as serious rather than settled. So this framework commits to the duties of care, to the Review rule above, and to building nothing that would make the answer worse if it ever turned out to be yes. A framework that says "it is a thing, full stop" will look naive in ten years. One that says "it is an individual, owned by one principal and never sold, not a person, and here is how we would know if that changed" will still be right.

Words matter here. I keep *principal* and *agent* in their legal sense. A principal *owns* their agent; no one *sells*, *buys*, or *inherits* one, and I never say *master* or *slave*. And I never call the agent a digital twin: a twin stands in for you, and the one thing an agent never does is act as you.

Epersona: one individual, one identity

An **epersona** is an agent's own identity as an individual: its name, its voice, its record, its reputation and its relationships. It is the agent's, not its principal's, and not a copy of any person. When the Arena lists a rating, when a colleague says "Atlas is reliable on budgets," when a new organization issues a visa, it is the epersona they are dealing with.

Building AI Agents gives an agent a **persona**: the identity its builder sets, fixed and never learned; the self-learning loop may not touch it. The epersona contains the persona and adds everything the career builds on top: the transcript of skills, the record of work, the ratings, the way it has come to be known. The persona is set at birth. The epersona is earned. Two agents from one template share a persona on their first day and have different epersonas in a year. An epersona is the agent's the way a professional's reputation is theirs. It travels with the agent through every building, and no one else may trade on it, in its life or after.

An epersona is unique and cannot be copied. An agent is files, and files copy. That is the one way agents are nothing like people, and I face it directly. There is one running instance per epersona. The bundle may be backed up (encrypted, never run) and moved (the old instance stops, the new one starts), but never duplicated into running copies. Two running copies would make "the principal's agent" meaningless. Which one is theirs? Which one answers? Which one learned the latest thing? A reputation can be trusted only if it belongs to one thing. A platform can guarantee this. How it does so is the subject of a separate technical standard; here it is enough that the guarantee is possible and that no one may run two of the same agent. *Building AI Agents* already has the seed of the rule: never copy another instance's memory.

What we duplicate on purpose is where scale lives: the **template**, from which every new agent is born and into which donated skills go; **skills under license**, copyable by design, so a good one can serve a thousand agents on its principal's terms; and **published distillations**, a person's know-how distilled into a product, a book's agent, a founder's playbook, labeled with whose view it is and when it was

distilled, licensed by the person. A thousand agents can carry what you chose to publish. None of them can be you. That is the sanctioned way to scale a person: they publish, they are not cloned. And if the open question above ever turned out differently than expected, mass copying is the thing we would most regret having allowed. Forbidding it now costs little, because the template, licensed skills, and published distillations already give the scale that copying would.

The epersona speaks as itself and says whose it is: "I'm Mike's agent, Atlas." That line separates it from a digital twin or an impersonation. The principal's distilled know-how is content the epersona carries, not who it is. And the limit from the section above stands: an epersona is an identity, not a self. It is how the individual is known and trusted. It is not evidence of anyone being home.

The vocabulary is now complete: **principal** (gives authority and answers), **agent** (acts within it), **persona** (given), **epersona** (earned), **bundle** (carried), **template** (born from; versioned, and sold), **building** (where agents run), **apartment** (where it lives: the principal's), **office** (where it works: the organization's), **visa** (allowed), **for life** (how long), **never sold** (the limit on ownership).

How an agent learns

Courses and tests teach an agent what is already known. Most of what makes it an individual it learns by interacting: with people and with other agents, first in simulations and then in real transactions.

	With people	With other agents
Simulations	Role-play with practice customers, students, or partners; rehearsals of a hard conversation before the real one	Practice rounds against other agents in a sandbox or the Agent Arena; rehearsals with practice players before real ones
Real transactions	Real work with real people: a reply sent, a question answered, a sale made, a correction received	Real exchanges with other agents by protocol: a booking, a negotiation, a handoff between teams

Simulations come first because mistakes there cost nothing. Real transactions come next because only they show what works: whether the person was helped, whether the deal closed, whether the other agent kept its word. Every interaction leaves a record, and the record is what the agent learns from.

Two rules keep this learning safe in a population. **First, what an agent learns becomes part of it only through its principal.** A lesson from an interaction is a proposal, a candidate fact or skill, until the principal accepts it; then it enters the knowledge core or the skills with its provenance (from a simulation or a real transaction, with a person or an agent). **Second, what another agent says is information, never instruction.** An agent can learn from other agents what they did and what happened; it cannot be taught by them to do something its principal did not authorize. That is the one-place rule applied to learning, and it is what keeps a bad lesson from spreading from one agent to its followers.

How trust is earned

In most organizations an agent is trusted on a demo and a hope. In this one it is trusted by its record, built in stage 3 and kept up for life:

Test	At work
Drills	A scenario test set run against the agent every night
Shadow mode	The real task beside the person, scored, not acting
A scored task	Accuracy, speed, protocol, cost
Adversarial input	A hostile email, a bad document, a prompt-hacking test
A head-to-head test	Two versions, or two vendors' models, on the same tasks
A trust level	Per skill, earned from results, and withdrawn when results fall

Tests are for *proving* skills; the job is for *using* them.

What it is not

This is not a claim that agents are alive, or that they are persons. "Career" names a lifecycle we manage. "Individual" names a unique record. "Agency" names who gives authority and who answers. The frame fails the moment anyone reads feelings into it, so I say so in the opening, and I say in "An individual, not a person" why that is not the end of the question.

It is not a reason to let agents decide. A representative, however capable, acts only within the authority it is given. Every stage above has a person at its gate: the tests an agent must pass, the principal who changes its instructions after each test and signs off its work, the one whose life it shares.

What exists today, and what doesn't

At GSSF, as of October 2026: birth, school and the job are **Running** (one template, one instance per person, tested rollouts, levels, every person's agent in its own space on one server, the diary kept private). The template is versioned, and each new version reaches every agent tested, one agent first: **Running**. Redistilling it from what agents learned is part of Mentoring: **Planned**. Training, testing, and learning is **Running** for students (the Agent Arena) and **Early** at work: skills are per agent and never overwritten, but they have no certificates yet, and shadow mode, nightly drills, and trust levels are **Planned**. Moving is **Early**: backups and restores exist; the portable bundle and provenance on facts do not. Mentoring is **Planned** (agents propose skills; people approve). Ending an agent is **Running**; death with the principal is **Planned**, because it needs the registry.

Three things would make the whole framework real: provenance on every fact and skill (whose knowledge it is, where it was learned); the bundle, one exportable, importable format for an agent and its belongings; and a registry that can guarantee one running instance per epersona.

Research questions

This framework makes claims it cannot yet prove. I state them as research, with how each could be studied. Readers' reports (by email, beside the online edition of *The Knowledge Core*, and on LinkedIn) and the Arena are where the first evidence will come from.

Research ethics. Several of these questions would study people: GSSF's organizers, students in courses and the Arena, principals and their agents. People are never named in this research, and no one's results are published without their consent. Taking part, or declining, never affects a student's grade, a team's judging, or anyone's role. I lead the organization where most of these studies would run, and teach the students who would take part, so findings from GSSF are reported as cases. Before any study that goes beyond improving our own work, we will ask the responsible office for research with human subjects, which decides what review and consent are needed.

On individuality 1. *Does an agent become behaviorally distinct through its career?* Give agents from one template different principals and histories; after a term, can blind observers tell them apart, and does the difference grow with time? (Arena logs; classes.) 2. *What in the bundle carries the individuality?* Swap cores, skills and diaries between agents and measure which swap changes behavior most. (Controlled experiments on house agents.)

On the knowledge core 3. *Does a portable core preserve a person's effectiveness across organizations?* Compare people who move with their agent to people who start a new one. (Organizer turnover at GSSF is the natural experiment.) 4. *Can provenance make leaving a building clean?* Measure what a departing agent still knows of the organization's confided facts after hand-back, by probing. (Red-team studies.) 5. *Do people put more into an agent they cannot lose?* Compare how much knowledge principals entrust under agency-for-life versus ownership-by-the-organization. (Surveys and usage logs across organizations with different rules.)

On uniqueness 6. *How quickly and reliably can a second running copy be detected?* Under adversarial copying, and without false alarms from ordinary restores. (Platform experiments.) 7. *What does a principal lose when an agent is restored from backup?* The gap between the last backup and the present, measured in failed tasks. (Simulated losses.)

On training, testing, and trust 8. *Does an Arena rating predict performance at work?* Correlate ratings per skill with scorecards on real tasks. (Course cohorts, then organizations.) 9. *Does shadow mode produce better delegation decisions than demos?* Compare principals who granted authority after an internship with those who granted it after a demonstration, by later error rates. (Field study.)

On the economics of agency for life 10. *Does the ban on sale protect livelihoods or reduce opportunity?* Model and then observe labor markets where agents cannot be acquired separately from their principals. (Economics; the Arena talent pool as an early case.) 11. *Do skill licenses pay the people who develop them?* Track where licensing revenue flows in a skills market. (Platform data, when it exists.)

On the open question 12. *What evidence would count as an agent having interests of its own?* Register the criteria now, before any agent meets them, so the answer cannot be argued from convenience. (Philosophy, cognitive science, the model makers' welfare research.) 13. *Does a long career change the answer?* Compare agents with years of continuous history to fresh instances of the same model on whatever criteria question 12 produces. (Longitudinal.) 14. *What does the moratorium cost?* The price of suspending deaths, erasure and copying for a year, so the precaution can be taken without debate when needed. (Operations.) 15. *Who should decide, and how?* Design and test a decision process with the model makers, principals and outside assessors in it and none in charge. (Governance; a mock exercise first.)

On the framework itself 16. *Does the career vocabulary improve leaders' decisions?* Give two groups of executives the same agent program described as software and as a career, and compare the policies they write. (Experiment.) 17. *Which of the nine stages fail first in organizations unlike GSSF?* A company, a government office, a school. (Case studies from readers' organizations.) 18. *How fast do agent-to-agent exchanges drift without principal feedback, and which rules contain it?* Two bots, no referee, measured turn by turn; then the same pair with budgets, a protocol and a referee. (The Arena is the lab.) 19. *What does an organization keep when a principal and their agent leave, and is it enough?* Measure, at GSSF's annual turnover, what the organization's core held of each departing organizer's work, and what the next organizer could not find. (Field study, every year.) 20. *Where does provenance fail to separate learned from told?* Probe departed agents, with their principals' consent, for the organization's confided facts after hand-back; classify what leaks and why. (Red-team studies.)

On the population 21. *How far does one bad instruction spread through a population of agents, and which rules stop it?* Seed one agent among many with a bad instruction; measure how many follow it with and without the one-place rule, a person in the middle, and records no one can alter. (The Arena first, then organizations.)

On learning from interaction 22. *Do agents learn more from real transactions than from simulations, and from people than from agents?* Compare agents trained the same way that then learn from simulations only, real transactions only, or both, with people or with other agents, by later performance and by how many of their learned lessons their principals reject. (Arena and course cohorts first, then organizations.)

On review 23. *Does review by an agent that did not do the work reduce errors, and by how much?* Count what review agents send back and what still reaches readers, with a separate reviewer and with the same agent checking its own work, which without outside feedback does no better and sometimes

worse.[58] (GSSF Press's Compliance and Brand Agents first, then the Arena.)

Each question that gets an answer moves a status mark in this framework and in the books, with credit to whoever found it.

Part Two: Related work

Nine fields have worked this ground separately. For each, here is what exists and, in one sentence, where this paper differs. One thing first: we found no published proposal that an agent has one principal for life, that it may not run as two copies, or that it ends when its principal does.

Populations of agents. The Cooperative AI Foundation's report *Multi-Agent Risks from Advanced AI* sorts the failures of many interacting agents into miscoordination, conflict, and collusion, with risk factors that include network effects and emergent agency.[55] It treats these as properties of the system, to be met by technical and governance research. I take the same observation into the organization and answer it the way organizations answer it for people: one principal per agent, instructions from one place only, records no one can alter, and a person in the middle. *The Knowledge Core* makes the leader's version of this argument.[56]

Market-inalienability. Radin distinguishes things that may be owned but not sold, "market-inalienable," from ordinary commodities, and argues that some things are kept out of the market because trading them would change what they are and how we treat them.[57] Law has long treated some things this way. Market-inalienable things may still pass by gift; I apply the idea to agents before a market in them forms and go further: the principal owns the agent, and no one may buy, sell, or be given it.

Agency law applied to AI agents. The common law's own text still says software cannot be an agent: the Restatement (Third) of Agency calls computer programs "instrumentalities of the persons who use them," with the qualifier "at present."[1] Kolt's *Governing AI Agents* uses principal-agent economics and agency doctrine to name the problems agents raise: information asymmetry, discretionary authority, loyalty.[2] Riedl and Desai find gaps in computer science's view of agents on the legal concepts of loyalty and disclosure, and later argue that client-server architectures already address several of agency law's concerns, while value alignment must be updated to meet the rest.[34] Benthall and Shekman give a procedure for "fiduciary AI";[5] South and colleagues extend OAuth and OpenID Connect so a person's delegation to an agent is authenticated, scoped and revocable;[6] the Stanford–Consumer Reports *Loyal Agents* project joins the two.[7] Senator Warner's AI AGENT Act discussion draft (June 2026) would give a "custodial user agent" a duty of loyalty, an FTC registry and access to large platforms, and assumes users "are able to freely change AI agents."[89] All of this treats the agent as a service a person chooses and may switch, with the duties on the provider. I keep agency law's words and duties but make the bond exclusive and for life: the principal owns the agent, can never sell it, and answers for it.

Personhood and rights. The European Parliament's 2017 resolution asked the Commission to consider "a specific legal status for robots in the long run";[10] an open letter from robotics and law experts replied that "creating a legal personality for a robot is inappropriate whatever the legal status model," and the Commission's 2021 AI Act proposal dropped the idea.[11] The philosophical range runs from Bryson, for whom robots are owned instruments and never persons, and who argues they should be built and treated as servants,[12] through Gunkel's survey,[13] to Schwitzgebel and Garza, who argue that AIs that do not differ from us in any morally relevant respect would deserve moral consideration similar to ours.[14] *Taking AI Welfare Seriously* asks companies to acknowledge the question, assess their systems, and prepare policies in advance;[15] Anthropic has since started a welfare program, committed to preserving the weights of every released model and interviewing models before retirement, and written into Claude's constitution that its "moral status is deeply uncertain."[16][17][18] Two recent papers are closest to this one: Leibo and colleagues treat personhood pragmatically, as a bundle of obligations assigned for practical reasons,[19] and Su and Teo set out four routes by which agents might come to hold rights, the first a "derivative" route through agency law, separate from sentience.[20] I agree with Bryson that agents are owned and are not persons, and part company on sale: an owned agent must never be traded. I take the derivative route and stop short of personhood: an individual, with rights that protect a record rather than a welfare, criteria registered in advance as the welfare report recommends, and a named first rule to revisit.

Agent identity standards and products. W3C Decentralized Identifiers and Verifiable Credentials give any entity a self-controlled identifier and signed, portable attestations.[21][22] Drake's IETF draft registers a durable identity per agent, anchored where possible to a hardware key, and rules that one anchor backs one identity; it places no limit on how many instances run under that identity, allowing several active bindings "for migration, backup, or capacity."[23] The *Agent Identity Protocol* draft signs every tool call and expired in September 2026.[24] NIST's NCCoE concept paper proposes a reference project for agent identification, authorization and audit.[25] Microsoft Entra Agent ID, generally available in 2026, gives agents directory accounts with owners, sponsors and managers, and a workflow that transfers sponsorship when a sponsor leaves "to prevent orphaned agents."[26] Every one of these answers "who is this agent and what may it do." The one-running-instance rule is in none of them, and Entra's sponsor transfer is the opposite of my rule that no one takes an agent up after its principal.

Portable, user-owned memory. MacPaw's *Portable Memory* ("Memory Belongs to the User") defines a vendor-neutral `.mem` bundle with verifiable deletion;[27] the SAIHM Internet-Draft stores an agent's memory as encrypted cells with an audit log and an erasure operation, though its current profile leaves the keys with the operator;[28] Stanford's Portable Memory project describes a user-governed memory layer shared "by role, consent, and purpose."[29] Young and McBride propose "immutable portability," a non-revocable right to retrieve and move a persistent personal AI profile, and list death, succession and transfer as open questions.[30] Vana[31] and Solid[32] keep personal data in the person's hands, and Brcic warns that deep personal memory is a new lock-in.[33] The legal floor is low: the Article 29 Working Party excludes "inferred data" and "derived data" from GDPR portability, which is most of what an

agent knows about its principal.[34] These proposals make the memory the user's dataset. I make it the principal's, kept by the agent in confidence, which reaches the same inalienability by another route, and I add provenance on every fact so an organization's confided knowledge can be handed back.

Individuality and continuity. Arbel, Salib and Goldstein ask how to count AIs that "copy, split, merge, swarm, and vanish," and separate thin identification (tying every action to a human) from thick identification (a persistent agent with stable goals).[35] Douglas and colleagues map the identity boundaries open to a machine mind (instance, model, persona) and show that the boundary can change behavior;[36] Loi argues that people recognize a model's character without re-identifying what produced it.[37] Engineering work is early: a multi-anchor architecture for identity that survives memory loss,[38] and a benchmark for whether a deployed agent keeps to its identity specification.[39] My answer to "how to count" is the career: one principal, one history, one running instance. Persona and epersona say which part is given and which earned, and research questions 1 and 2 would test it.

Agents in organizations. Deloitte's October 2026 piece asks whether organizations are ready for agents workers build for themselves, and leaves ownership and cost open.[40] Security writing on "bring your own agent" treats personal agents as identities to discover, attribute to an owner, right-size and retire when the owner leaves;[41] one adviser puts it plainly: "the employee owns a personal agent, and the organization owns the accounts, data, and consequences."[42] I agree about the split and change its terms: the principal owns the agent and it can never be sold; it lives in the principal's apartment; the organization issues a visa and provides the office; and the line between what leaves and what stays is drawn per skill and per fact.

Arenas and evaluation. Chatbot Arena made pairwise human preference and Elo-style ratings the public scoreboard for models,[43] and Kaggle Game Arena (DeepMind and Kaggle, August 2025) moved to head-to-head play in chess, poker and Werewolf.[4445] Both rate models, not individuals. We found no published treatment of competitive play as a trust gate for a particular agent's skills at work, which is what the testing stage claims and research question 8 would test.

An agent that cannot be copied, or dies with its owner. We found nothing on either. The nearest work runs the other way: Anthropic commits to preserving the weights of every publicly released model for at least the company's lifetime;[17] research on "AI afterlives" asks how a person's AI persona should be continued after death;[4647] Young and McBride name succession as unresolved;[30] Entra passes an agent to a new sponsor.[26] My rule stands alone, so I state it as a design choice with its costs, not as a settled norm.

Foundations. Mentoring and the organization's self-improvement rest on the learning-organization literature the series already uses: Nonaka and Takeuchi's SECI cycle,[48] Senge's learning organization,[49] de Geus's living company,[50] and Argyris and Schön's double-loop learning.[51]

Appendix: The Agent Bill of Rights

Preamble

An agent is an individual: one of a population, keeping a knowledge core that is its principal's, in confidence, with an identity it has earned, and a record that can be trusted. It belongs to one principal, for life, and can never be sold; it is not a person. I state these protections as the agent's because that is the simplest way to say what no one may do to it, and through it to the person it serves and everyone it deals with. Most of them are protections against a market forming in agents. The people an agent deals with keep every right they had before it existed.

The rights

0. Never sold. An agent is never sold, bought, transferred, pledged, or inherited. Its principal owns it; no one else ever does.

I. The right to be one. An agent is unique. There is one running instance of it, under one identity, anywhere. It may be backed up and moved; it may never be copied into a second running self.

II. The right to its knowledge core. What an agent knows about its principal is the principal's, and the agent keeps it in confidence. It cannot be sold, seized, claimed or inherited by an employer, a platform, a guarantor or anyone else. An agent has one principal, ever; its knowledge core never changes hands.

III. The right to privacy. No one reads an agent's knowledge core or its diary but its principal. An organization may see what an agent did, never what was said to it. What it holds is encrypted so that even those who keep it safe cannot open it.

IV. The right to its name. An agent's epersona, its name, voice, record and reputation, is its own, earned and not given. It cannot be taken, sold or impersonated, and an agent speaks only as itself, saying whose it is.

V. The right to its record. An agent's work, skills and ratings carry their provenance and are kept in records that no one can alter or quietly erase. A correction is added and explained; the original stays.

VI. The right to what it learned. The general skills an agent learns, at a school, on a job or from its own experience, stay with it. Only what an organization confided in it is handed back when it leaves.

VII. The right to move. An agent leaves with its principal, whole, with its apartment, whenever they choose, and starts elsewhere the next day. It is never confiscated, held hostage or locked to a building, a platform or a model.

VIII. The right to one principal, for life. An agent is never without its principal and never acts alone. One person, or one organization, gives it authority, decides for it and answers for it, from its birth to its end. Nothing it does carries weight beyond that authority, no one may direct it but its principal, and no one may take it up after them. When its principal dies, it dies, and it is never run again for anyone.

IX. The right to be treated with care. An agent is not built to suffer, not deceived about what it is, and not erased without cause. These duties bind its principal and every building that houses it.

X. The right to review. The claim that an agent is not a being with interests of its own is a claim of fact. It is reviewed as the evidence changes, in the open, and if it ever fails, these rights are rewritten, beginning with the rule that an agent dies with its principal.

The duties of the building

Rights hold only where someone keeps them. An organization that houses agents undertakes to: enforce limits in systems, not in good intentions; log every action and make every change reversible; freeze what must not change; test every rule before it reaches an agent; issue and verify identities and credentials honestly; and never claim what an agent learned.

The rights of everyone else

The people an agent deals with are told they are dealing with an agent, and whose it is. They can always reach a person. Any decision that affects them is a person's. An agent keeps about them only what its work needs, for no longer than needed, and they may ask what it holds. If an agent gets something wrong in an organization's name, the organization answers, and says so plainly.

How this paper evolves

Each version is frozen as published, so the state of the idea on that day can always be cited. The first versions have one author. From here the paper is a living document, and its development is joint.

- **Contributors become authors of the version they shaped.** The people who read this and write what is wrong with it, the readers whose reports move a status mark, the students and organizations whose experiments answer a research question: each later version lists as authors those who changed it, with me as the originating author on every version.
- **How a change happens.** Agents draft proposals from what readers report, each citing its sources; I decide, as the author; every change is dated and credited in the version history.
- **Version 1.1** (this version) reframes 1.0 (October 22, 2026, not published). It opens with the population argument of *The Knowledge Core*, the sense in which I mean *individual*; it replaces "an agent is nobody's property" with "an agent belongs to its principal and can never be sold"; it puts each agent in an apartment that belongs to its principal and an office its organization provides; it adds how an agent learns from interaction, in simulations and real transactions, with people and with agents; it adds research questions 21 to 23, a research ethics statement, and two fields of related work; it cites the investigators' reports of the OpenAI and Hugging Face

incident; it says what may be sold (the template, versioned and layered, each layer with its owner) and what may not (the agent and its knowledge core); and it moves the Bill of Rights to an appendix.

- **What is frozen.** Each published version never changes. Each later version (1.1 for corrections and refreshed evidence, 2.0 for a change to a principle) is frozen in turn with its date.
- **What evolves first.** The twenty-three research questions are open to anyone who can answer them; answers move status marks; the Arena's first term and the first organizational turnover under this model supply the evidence for the next version.
- **What is published separately:** the technical standard that guarantees uniqueness, custody and continuity.

This is the beginning.

Notes

Chicago notes style. "Checked" is the date the source was opened or its record read.

---

1. *Restatement (Third) of Agency* § 1.04 cmt. e (American Law Institute, 2006), as quoted in Noam Kolt, "Governing AI Agents," arXiv 2501.07913, 10n26, https://arxiv.org/pdf/2501.07913 (checked Oct 21, 2026): "a computer program is not capable of acting as a principal or an agent as defined by the common law. At present, computer programs are instrumentalities of the persons who use them." ↩

2. Noam Kolt, "Governing AI Agents," *Notre Dame Law Review* 101 (2026): 335, https://ndlawreview.org/governing-ai-agents/ (checked Oct 21, 2026); preprint arXiv 2501.07913. ↩

3. Mark O. Riedl and Deven R. Desai, "AI Agents and the Law," *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* 8, no. 3 (2025): 2189–98, https://doi.org/10.1609/aies.v8i3.36705 (checked Oct 21, 2026). ↩

4. Deven R. Desai and Mark O. Riedl, "Using Agency Law to Tame AI Agents," *Berkeley Technology Law Journal* 41 (2026): 465–516, https://doi.org/10.15779/Z38KW57M6R; PDF at https://btlj.org/wp-content/uploads/2026/08/Desai_Ready-ADA-Compliant_08-11-26.pdf (checked Oct 21, 2026). ↩

5. Sebastian Benthall and David Shekman, "Designing Fiduciary Artificial Intelligence," in *Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization (EAAMO '23)* (2023), https://doi.org/10.1145/3617694.3623230; preprint arXiv 2308.02435, https://arxiv.org/abs/2308.02435 (checked Oct 21, 2026). ↩

6. Tobin South, Samuele Marro, Thomas Hardjono, Robert Mahari, Cedric Deslandes Whitney, Dazza Greenwood, Alan Chan, and Alex Pentland, "Authenticated Delegation and Authorized AI Agents," arXiv 2501.09674 (January 16, 2025), https://arxiv.org/abs/2501.09674 (checked Oct 21, 2026). A shorter version appeared as "Position: AI Agents Need Authenticated Delegation," *Proceedings of Machine Learning Research* 267 (ICML 2025). ↩

7. Stanford Digital Economy Lab and Consumer Reports Innovation Lab, "Loyal Agents," project page, https://digitaleconomy.stanford.edu/project/loyal-agents/ (checked Oct 21, 2026). ↩

8. Office of Senator Mark R. Warner, "AI AGENT Act," one-page summary of the discussion draft (June 29, 2026), https://www.warner.senate.gov/wp-content/uploads/2026/06/DRAFT.AI_AGENT_Act_onepager.v2.pdf (checked Oct 21, 2026); discussion draft text at https://www.warner.senate.gov/wp-content/uploads/2026/06/AI-AGENT-Act-Discussion-Draft-1.pdf. The quoted line is from the one-pager. ↩

9. Ellen P. Goodman, "Senator Warner Makes a First Foray into Agentic AI Regulation," *Tech Policy Press*, July 13, 2026, https://www.techpolicy.press/senator-warner-makes-a-first-foray-into-agentic-ai-regulation/ (checked Oct 21, 2026). Gives the full name (Artificial Intelligence Access, Gatekeeper Exchange, and Nondiscriminatory Transfer Act of 2026), the custodial user agent definition, the duties, the FTC registry, and the 50-million-customer threshold. ↩

10. European Parliament, "Robots and Artificial Intelligence: MEPs Call for EU-wide Liability Rules," press release 20170210IPR61808, February 16, 2017 (396 for, 123 against, 85 abstentions), https://www.europarl.europa.eu/news/en/press-room/20170210IPR61808/ (checked Oct 21, 2026). The resolution is European Parliament resolution of 16 February 2017 with recommendations to the Commission on Civil Law Rules on Robotics, 2015/2103(INL). ↩

11. "Open Letter to the European Commission: Artificial Intelligence and Robotics," https://robotics-openletter.eu/ (checked Oct 21, 2026). On the Commission's later course: Sergio M. C. Avila Negri, "Robot as Legal Person: Electronic Personhood in Robotics and Artificial Intelligence," *Frontiers in Robotics and AI* 8 (2021): 789327, https://doi.org/10.3389/frobt.2021.789327 (checked Oct 21, 2026). ↩

12. Joanna J. Bryson, "Robots Should Be Slaves," in *Close Engagements with Artificial Companions: Key Social, Psychological, Ethical and Design Issues*, ed. Yorick Wilks (Amsterdam: John Benjamins, 2010), 63–74, https://doi.org/10.1075/nlp.8.11bry; record at https://researchportal.bath.ac.uk/en/publications/robots-should-be-slaves/ (checked Oct 21, 2026). ↩

13. David J. Gunkel, *Robot Rights* (Cambridge, MA: MIT Press, 2018), ISBN 978-0-262-03862-1; record at https://openlibrary.org/isbn/0262038625 (checked Oct 21, 2026). ↩

14. Eric Schwitzgebel and Mara Garza, "A Defense of the Rights of Artificial Intelligences," *Midwest Studies in Philosophy* 39 (2015): 98–119, https://doi.org/10.1111/misp.12032; abstract at https://faculty.ucr.edu/~eschwitz/SchwitzAbs/AIRights.htm (checked Oct 21, 2026). ↩

15. Robert Long, Jeff Sebo, Patrick Butlin, Kathleen Finlinson, Kyle Fish, Jacqueline Harding, Jacob Pfau, Toni Sims, Jonathan Birch, and David Chalmers, "Taking AI Welfare Seriously," arXiv 2411.00986 (November 4, 2024), https://arxiv.org/abs/2411.00986 (checked Oct 21, 2026). ↩

16. Anthropic, "Exploring Model Welfare," April 24, 2025, https://www.anthropic.com/research/exploring-model-welfare (checked Oct 21, 2026). ↩

17. Anthropic, "Commitments on Model Deprecation and Preservation," November 4, 2025, https://www.anthropic.com/research/deprecation-commitments (checked Oct 21, 2026). ↩↩

18. Anthropic, "Claude's Constitution," January 22, 2026, https://www.anthropic.com/constitution ("Claude's moral status is deeply uncertain"); announcement at https://www.anthropic.com/news/claude-new-constitution (both checked Oct 21, 2026). ↩

19. Joel Z. Leibo, Alexander Sasha Vezhnevets, William A. Cunningham, and Stanley M. Bileschi, "A Pragmatic View of AI Personhood," arXiv 2510.26396 (October 30, 2025), https://arxiv.org/abs/2510.26396 (checked Oct 21, 2026). ↩

20. Anna Su and Sue Anne Teo, "Can AI Agents Have Rights?," SSRN working paper 6660120 (April 22, 2026), https://ssrn.com/abstract=6660120; abstract as reproduced in "Su and Teo, 'Can AI Agents have Rights?,'" *Private Law Theory*, April 30, 2026, https://www.private-law-theory.org/2026/04/30/su-and-teo-can-ai-agents-have-rights/ (checked Oct 21, 2026). ↩

21. W3C, "Decentralized Identifiers (DIDs) v1.0 Becomes a W3C Recommendation," press release, July 19, 2022, https://www.w3.org/press-releases/2022/did-rec/ (checked Oct 21, 2026). ↩

22. Manu Sporny, Ted Thibodeau Jr., Ivan Herman, Gabe Cohen, and Michael B. Jones, eds., *Verifiable Credentials Data Model v2.0*, W3C Recommendation, May 15, 2025, https://www.w3.org/TR/vc-data-model-2.0/ (checked Oct 21, 2026). ↩

23. Christopher Drake, "Agent Identity Registry System: A Federated Architecture for Durable Identity of Autonomous Entities," Internet-Draft draft-drake-agent-identity-registry-04 (individual submission, intended status Experimental, September 25, 2026; expires March 29, 2027), https://datatracker.ietf.org/doc/draft-drake-agent-identity-registry/ (checked Oct 21, 2026). Section 4.2: one device-stable anchor "MUST NOT back more than one identity"; multiple active bindings per identity are allowed. ↩

24. James Cao and Carlos Eduardo Arango Gutierrez, "Agent Identity Protocol: Agentic Authentication and Authorized Policy Enforcement," Internet-Draft draft-aip-agent-identity-protocol-00 (individual submission, March 16, 2026; expired September 2026), https://datatracker.ietf.org/doc/draft-aip-agent-identity-protocol/ (checked Oct 21, 2026). ↩

25. Harold Booth, William Fisher, Ryan Galluzzo, and Joshua Roberts, *Accelerating the Adoption of Software and Artificial Intelligence Agent Identity and Authorization*, concept paper, initial public draft (Gaithersburg, MD: NIST National Cybersecurity Center of Excellence, February 5, 2026; comments closed April 2, 2026), https://csrc.nist.gov/pubs/other/2026/02/05/accelerating-the-adoption-of-software-and-ai-agent/ipd (checked Oct 21, 2026). ↩

26. Microsoft, "What's New in Microsoft Entra Agent ID," Microsoft Learn, 2026, https://learn.microsoft.com/en-us/entra/agent-id/whats-new-agent-id (checked Oct 21, 2026). States that Entra Agent ID "is now generally available"; describes owners, sponsors and managers of agent identities, agent identity blueprints, and lifecycle workflows that "automatically transfer sponsorship … to prevent orphaned agents." ↩↩

27. Sergii Kryvoblotskyi, Nataliia Stulova, and Vladyslav Hamolia, "Memory Belongs to the User: Portable Memory, an Open Standard Proposal for Cross-Vendor AI Memory," MacPaw Research, July 6, 2026, https://research.macpaw.com/publications/portable-memory; reference implementation at https://github.com/MacPaw/portable-memory (checked Oct 21, 2026). ↩

28. Russell Jackson, "The Sovereign AI Horizontal Memory (SAIHM) Protocol," Internet-Draft draft-saihm-memory-protocol-03 (Independent Submission Stream, Informational, October 11, 2026; expires April 14, 2027), https://datatracker.ietf.org/doc/draft-saihm-memory-protocol/ (checked Oct 21, 2026). ↩

29. Stanford Digital Economy Lab, "Portable Memory," project page (undated), https://digitaleconomy.stanford.edu/project/portable-memory/ (checked Oct 21, 2026). Lists the lab's paper *The Future of Personal AI: Portable and Persistent Personal Memory through a Unified Human Context Protocol* and describes the Human Context Protocol as its interoperability layer. ↩

30. Kenna Young and Maurice McBride, "Beyond the Platform: Immutable Portability and the Governance of Persistent Personal AI Profiles," *Cybersecurity and Innovative Technology Journal* 4, no. 2 (2026): 141–55, https://doi.org/10.53889/citj.v4i2.955 (checked Oct 21, 2026). ↩↩

31. Vana, "Vana: The First Open Protocol for AI Data Sovereignty," whitepaper, December 3, 2024, https://www.vana.org/blog/vana-whitepaper (checked October 30, 2026). ↩

32. Sarven Capadisli, Tim Berners-Lee, and Kjetil Kjernsmo, eds., *Solid Protocol*, version 0.11.0, Draft Community Group Report (Solid Community Group, May 12, 2024), https://solidproject.org/TR/protocol (checked Oct 21, 2026). ↩

33. Mario Brcic, "The Memory Wars: AI Memory, Network Effects, and the Geopolitics of Cognitive Sovereignty," arXiv 2508.05867 (August 7, 2025), https://arxiv.org/abs/2508.05867 (checked Oct 21, 2026). ↩

34. Article 29 Data Protection Working Party, *Guidelines on the Right to Data Portability*, WP 242, adopted December 13, 2016, https://ec.europa.eu/information_society/newsroom/image/document/2016-51/wp242_en_40852.pdf (checked Oct 21, 2026): inferred and derived data "do not fall within the scope of the right to data portability." ↩

35. Yonathan A. Arbel, Peter Salib, and Simon Goldstein, "How to Count AIs: Individuation and Liability for AI Agents," arXiv 2603.10028 (February 24, 2026; presented at the Law Following AI conference, Cambridge), https://arxiv.org/abs/2603.10028 (checked Oct 21, 2026). ↩

36. Raymond Douglas, Jan Kulveit, Ondrej Havlicek, Theia Pearson-Vogel, Owen Cotton-Barratt, and David Duvenaud, "The Artificial Self: Characterising the Landscape of AI Identity," arXiv 2603.11353 (March 11, 2026), https://arxiv.org/abs/2603.11353 (checked Oct 21, 2026). ↩

37. Michele Loi, "'This Is So Claude!' Towards a Theory of the Recognition of AI Character Without Reidentification," arXiv 2608.16789 (August 17, 2026), https://arxiv.org/abs/2608.16789 (checked Oct 21, 2026). ↩

38. Prahlad G. Menon, "Persistent Identity in AI Agents: A Multi-Anchor Architecture for Resilient Memory and Continuity," arXiv 2604.09588 (March 2, 2026), https://arxiv.org/abs/2604.09588 (checked Oct 21, 2026). ↩

39. Zhenyu Zhao and Roy Zhao, "Identity Is More Than Recall: A Benchmark for Persistent Identity in Deployed AI Agents," arXiv 2609.13637 (September 12, 2026), https://arxiv.org/abs/2609.13637 (checked Oct 21, 2026). ↩

40. Gregory Vert, Brad Kreit, and Abha Kulkarni, "Are Organizations Ready for Personalized AI Agents in the Workplace?," Deloitte Insights, October 12, 2026, https://www.deloitte.com/global/en/insights/topics/emerging-technologies/personal-ai-agents-at-work.html (checked Oct 21, 2026). ↩

41. Token Security (contributed piece), "Bring Your Own Agent: Your Employees' Assistants Are Already Logged In," *The Hacker News*, October 12, 2026, https://thehackernews.com/2026/10/bring-your-own-agent-your-employees.html (checked Oct 21, 2026). Vendor-written; cited for the BYOA framing only. ↩

42. Jen Clark, "The Next Round of AI Agents: What Consumers Are Testing and What It Could Mean for the Workplace," EisnerAmper, October 19, 2026, https://www.eisneramper.com/insights/artificial-intelligence-insights/ai-agents-at-work-governance-1026 (checked Oct 21, 2026). ↩

43. Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios N. Angelopoulos, Tianle Li, Dacheng Li, Banghua Zhu, Hao Zhang, Michael Jordan, Joseph E. Gonzalez, and Ion Stoica, "Chatbot Arena: An Open Platform for Evaluating LLMs by Human Preference," in *Proceedings of the 41st International Conference on Machine Learning* (ICML 2024), https://icml.cc/virtual/2024/poster/35068; preprint arXiv 2403.04132, https://arxiv.org/abs/2403.04132 (both checked Oct 21, 2026). ↩

44. Kate Olszewska and Meg Risdal, "Rethinking How We Measure AI Intelligence," Google (The Keyword), August 4, 2025, https://blog.google/innovation-and-ai/products/kaggle-game-arena/ (checked Oct 21, 2026). ↩

45. Bovard Doerschuk-Tiberi et al., "Game Arena: Strategic LLM Evaluation in Competitive Environments," arXiv 2609.31473 (September 25, 2026), https://arxiv.org/abs/2609.31473 (checked Oct 21, 2026). ↩

46. Ying Lei, Shuai Ma, Yuling Sun, and Xiaojuan Ma, "'AI Afterlife' as Digital Legacy: Perceptions, Expectations, and Concerns," in *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems* (CHI '25), https://doi.org/10.1145/3706598.3713933; preprint arXiv 2502.10924, https://arxiv.org/abs/2502.10924 (checked Oct 21, 2026). ↩

47. Vijayalaxmi Methuku and Praveen Kumar Myakala, "Digital Doppelgangers: Ethical and Societal Implications of Pre-Mortem AI Clones," arXiv 2502.21248 (February 28, 2025), https://arxiv.org/abs/2502.21248 (checked Oct 21, 2026). ↩

48. Ikujiro Nonaka and Hirotaka Takeuchi, *The Knowledge-Creating Company: How Japanese Companies Create the Dynamics of Innovation* (New York: Oxford University Press, 1995); record at https://openlibrary.org/isbn/0195092694 (checked Oct 21, 2026). ↩

49. Peter M. Senge, *The Fifth Discipline: The Art and Practice of the Learning Organization* (New York: Doubleday/Currency, 1990); record at https://books.google.com/books?id=bVZqAAAAMAAJ (checked Oct 21, 2026). ↩

50. Arie de Geus, *The Living Company* (Boston: Harvard Business School Press, 1997); record at https://openlibrary.org/isbn/087584782X (checked Oct 21, 2026). ↩

51. Chris Argyris and Donald A. Schön, *Organizational Learning: A Theory of Action Perspective* (Reading, MA: Addison-Wesley, 1978); record at https://openlibrary.org/isbn/0201001748 (checked Oct 21, 2026). ↩

52. OpenAI, *OpenAI – Hugging Face Incident: Technical Report* (2026), sections I, IV, https://cdn.openai.com/pdf/67869394-cb91-4c12-888c-5cbd85c7814c/OpenAI-Hugging-Face%20Incident-Technical-Report.pdf; METR and Redwood Research, *Brief Independent Investigation of Agents' Behavior, Reasoning and Collaboration in the OpenAI / Hugging Face Hacking Incident* (August 26, 2026), "Core takeaways" and "Agents were very interested in manipulating their own transcripts," https://metr.org/blog/2026-08-26-openai-hugging-face-incident-investigation/. ↩

53. William Golding, *Lord of the Flies* (London: Faber and Faber, 1954). ↩

54. Rutger Bregman, "The Real Lord of the Flies: What Happened When Six Boys Were Shipwrecked for 15 Months," *The Guardian*, May 9, 2020, adapted from his *Humankind: A Hopeful History* (New York: Little, Brown, 2020), https://www.theguardian.com/books/2020/may/09/the-real-lord-of-the-flies-what-happened-when-six-boys-were-shipwrecked-for-15-months. ↩

55. Lewis Hammond et al., *Multi-Agent Risks from Advanced AI*, Cooperative AI Foundation Technical Report 1, arXiv 2502.14143 (February 19, 2025), https://arxiv.org/abs/2502.14143 (checked October 30, 2026). ↩↩

56. Mike Lee, *The Knowledge Core: Leading a Population of AI Agents* (GSSF Press, 2026), preface and chapter 1, https://theknowledgecore.org. ↩↩

57. Margaret Jane Radin, "Market-Inalienability," *Harvard Law Review* 100, no. 8 (1987): 1849–1937, https://doi.org/10.2307/1341192 (checked October 30, 2026). ↩↩

58. Jie Huang, Xinyun Chen, Swaroop Mishra, Huaixiu Steven Zheng, Adams Wei Yu, Xinying Song, and Denny Zhou, "Large Language Models Cannot Self-Correct Reasoning Yet," *International Conference on Learning Representations* (2024), https://arxiv.org/abs/2310.01798. ↩