

Fall Forum 2025

Who Is the Human in the Loop?

Dignity, Judgment and Responsibility in the Agentic Workplace

Forum Summary and Rapporteur's Note

Digital Dignity Institute Fall Forum, Riviera Maya, Mexico

22 to 24 October 2025, Maroma, A Belmond Hotel

Thirty invited participants, convened under the Chatham House Rule

Executive Summary

The Forum convened on one question: what must remain true of a person at work for their involvement with an artificial agent to count as meaningful human judgment? By October 2025 the question had ceased to be speculative. Enterprise systems were being designed to plan, initiate, and complete work across ordinary software environments, while the safeguards attached to them still relied heavily on a familiar promise that a human would remain somewhere in the process. Participants found that promise too imprecise for the moral and organizational weight being placed upon it.

The central difficulty was not whether a worker appeared on a workflow diagram, but what that worker could actually know, decide, stop, revise, and refuse. A person may technically approve an action while lacking the time to examine it, the evidence to verify it, the expertise to challenge it, or the organizational latitude to slow it down. The Forum therefore distinguished human presence from human review, review from control, control from judgment, and judgment from authority. Those distinctions became the architecture of the meeting.

The sessions were deliberately practical. Participants worked through constructed workplace cases based on capabilities publicly available or announced by Salesforce, DeepL, Microsoft Research, OpenAI, and GitHub. The cases did not allege real incidents by those companies. They used real product affordances to force more exact questions: when an agent may act in the background, when a person can require approval, when an agent can learn and reuse a plan, when consequential actions require confirmation, and when a coding agent can complete a task and return a pull request for review, where precisely does the human decision reside?

Four findings structured the discussion. Human presence is not human control; oversight is a substantive form of work rather than a ceremonial click; autonomy moves by degrees and should be treated as an allocation of authority rather than a product setting alone; and responsibility is defensible only where meaningful authority remains with the person expected to bear it. The Forum added a fifth concern that grew in importance over three days: if organizations delegate the

very tasks through which junior workers acquire judgment, the long-term question is not simply what jobs remain, but where future human expertise will come from.

The Forum adopted the Maroma Principles for Dignified Human-Agent Work, ten propositions intended to make the phrase human in the loop materially more demanding. Their underlying proposition was simple enough to survive the three days of disagreement: the human in the loop is still a worker, with interests in authorship, professional standing, learning, voice, and the practical ability to exercise judgment. An organization cannot remove those conditions and then recover human responsibility by leaving an approval button behind.

This note reflects the rapporteur's impressions of the discussion. It is not a transcript, and no participant is committed in any way to its content or expression. In accordance with the Chatham House Rule, no view expressed here is attributed to any individual or institution, and the participant list in Annex A describes roles rather than names.

Suggested citation: Digital Dignity Institute, Who Is the Human in the Loop? Dignity, Judgment and Responsibility in the Agentic Workplace. Summary of the DDI Fall Forum, Riviera Maya, 22 to 24 October 2025.

The named company scenarios below are hypothetical workshop cases, not reports of failures or incidents at those companies. Their purpose is to anchor the discussion in the technological reality of the workplace landscape.

PART ONE

I. The loop

The Forum began by asking participants to define the term that had brought them together. The answers were sufficiently different that the first hour was spent not on policy but on vocabulary. A human could be described as in the loop because a system asked for approval before a consequential action, because a worker could interrupt an automated sequence, because a supervisor reviewed the final result, because a manager remained accountable for the process, or simply because an employee's name still appeared beside the completed work. Each arrangement contained a person. They did not contain the same kind of human authority.

This imprecision had become consequential during 2025 because the object under discussion was changing. Systems marketed as workplace assistants were acquiring the capacity to move through software, formulate multi-step plans, initiate actions in response to data, and continue working without a fresh human instruction at every stage. Salesforce described Agentforce 2dx as capable of proactive background action inside business processes without constant human oversight. Microsoft Research's Magentic-UI explored co-planning, co-tasking, action approvals, answer verification, memory, and the ability to alter how frequently approval was required. OpenAI's ChatGPT agent combined research with web action while requiring explicit confirmation for consequential acts and active supervision for some sensitive activities. DeepL Agent, then in beta with selected customers, emphasized the ability to monitor and manage tasks, stop and inspect an agent, and require human validation or approval. GitHub's coding agent allowed a developer to assign an issue to Copilot, let it work in the background, and receive a completed pull request for human review.

The technologies differed, sometimes substantially, but they exposed the same ambiguity. The phrase human in the loop described a location in a process while remaining largely silent about the quality of the person's participation. The Forum therefore chose to treat the workplace not as a two-party relationship between human and machine, but as a three-party arrangement among worker, artificial agent, and organization. The organization selects the system, configures its permissions, defines performance expectations, decides how much time may be spent reviewing outputs, and ultimately determines what happens when the worker disagrees. That third actor proved impossible to ignore.

The human can remain visible on the workflow long after the decision has migrated elsewhere.

From that point forward, participants were asked to identify five functions separately in every case: who set the objective, who determined the method, who performed the consequential action, who possessed the practical power to interrupt or alter the outcome, and who was expected to answer afterwards. In conventional work these functions are often already distributed. Agentic systems intensify the distribution while making it easier to speak as though one person still owns the whole decision.

PART TWO

II. The working sessions

Session One. What counts as human in the loop? Salesforce Agentforce and the location of judgment

The first session was designed to make the phrase human in the loop fail under ordinary use. Participants were given four versions of the same fictional customer-service workflow built around capabilities Salesforce had publicly described for Agentforce 2dx by March 2025, including proactive agents that could operate in the background of business processes and take action without constant human oversight. The fictional company was a national home-goods retailer whose customer-operations team handled delivery failures, replacements, credits, and post-purchase complaints through Salesforce. The issue was not whether Agentforce could be made to perform every detail exactly as written in the workshop case. The exercise used the platform's public 2025 direction as a realistic environment in which to test how quickly familiar language about human control becomes unstable.

In the first workflow, Agentforce assembled the customer's order history, extracted the relevant policy, proposed three possible responses, and drafted the correspondence. Nothing left the system until the employee selected an option, edited the text if necessary, and pressed send. Nearly everyone initially described the employee as the decision-maker, although several participants immediately observed that the system had already determined which facts were salient, which policy provisions mattered, and which response options appeared reasonable. The human retained the final act and considerable discretion, but the informational architecture of the decision had already been shaped upstream.

The second workflow moved one step further. The agent reviewed the complaint, recommended either replacement or refund, prepared the CRM entries, and placed a complete action in front of the worker for approval. The worker could reject it, but the company expected cases to be resolved in less than two minutes and the system had been accurate in the overwhelming majority of routine cases during the preceding month. Participants divided over whether this remained a human decision or had become a machine decision with human ratification. The distinction sounded scholastic until the rapporteur asked who should be blamed if the recommendation were wrong. At that point the room became markedly less relaxed about calling the worker the decision-maker.

The third workflow allowed the agent to resolve defined low-consequence matters on its own. Refunds below fifty dollars, replacement of inexpensive damaged items, and changes to delivery windows were processed automatically, while exceptions were escalated to a person. The worker's role shifted from deciding individual cases to supervising the boundary around the cases the agent was allowed to decide. Several participants argued that this was a perfectly coherent form of human control, but it was control at the level of policy rather than transaction. Others noted that the individual frontline employee might not have selected the threshold, written the policy, or possessed any authority to alter either one, which raised the question of whether the

relevant human in the loop was the worker, the team leader, the administrator who configured the system, or the executive who approved the deployment.

The fourth workflow removed routine cases almost entirely from view. Agentforce monitored incoming data, initiated the relevant workflow, communicated with customers, updated internal records, and sent only unusual or high-value situations to the service team. Workers received a daily summary and could inspect the history of a case after the fact. Participants were asked to place a marker on the exact point at which the human decision occurred. There was no consensus because there was no single point. Some located human judgment in the policy design that preceded the interaction; some in the exception rules; some in the later audit; and some concluded that ordinary cases no longer contained an individual human decision at all.

The result of the exercise was a taxonomy that the Forum used for the remainder of the meeting. Human presence meant that a person appeared somewhere in the workflow. Human review meant that a person saw an output before or after execution. Human control required a practical capacity to alter or stop the outcome. Human judgment required the person to form an independent assessment rather than merely acknowledge a recommendation. Human authority referred to the recognized power to determine what the system could do and when. The terms overlap, but they are not synonyms, and the session's most durable conclusion was that any governance regime that uses human in the loop without specifying which of these functions it means is likely to conceal more than it clarifies.

Session Two. The responsibility gap: a DeepL Agent case workshop

The second session began with a constructed case involving DeepL Agent, which in September 2025 was being tested in beta with selected customers and was publicly described as a general-purpose workplace agent capable of carrying out computer-based tasks while giving users and administrators the ability to monitor activity, stop and inspect the agent, and require human validation or approval. Participants were told to imagine a subscription-software company using such an agent inside its customer-support operation, where an employee supervised a stream of billing complaints that required movement between a support inbox, a browser-based billing system, an internal account panel, and the customer record.

A customer had written to request a forty-nine-euro service credit after an outage. The agent gathered the account history, confirmed that a credit was available under the company's policy, navigated the necessary systems, and prepared an approval card for the employee. Because of an error in interpreting a previous annual adjustment, however, it proposed a credit of four hundred and ninety euros. The company had configured the workflow so that credits below five hundred euros required one human approval but no managerial escalation. The employee, who had approved dozens of accurate cases that morning and had less than a minute allocated to each review, clicked approve. The credit was issued before the error was noticed.

The case was fictional, and participants were reminded of that point twice because the purpose of naming DeepL was not to imply that the company's system had produced such an error. The value of the scenario lay in the control architecture DeepL had actually made visible in 2025. A person could monitor the agent, stop it, inspect it, or require validation; administrators and managers could also exercise control over how the system was used. The workshop therefore asked a deliberately uncomfortable question: if the employee could technically have caught the error, but the employer selected the system, determined the threshold, designed the review queue, and set the productivity target, while the technology vendor built the agent, who should answer for the failure?

Participants were first asked to argue the case against the worker. The employee had seen the amount, possessed the power to reject it, and approved a plainly excessive credit. Several lawyers and managers regarded that as enough to establish at least some responsibility, particularly because human approval had been included precisely to catch cases in which the system was wrong. The counterargument was that the institution could not simultaneously describe the agent as reliable enough to multiply the employee's throughput, reduce the time available for review on that basis, and then describe the approval click as proof of a fully independent human judgment when the rare failure appeared.

The discussion then moved to the company. Participants asked who had chosen the five-hundred-euro threshold, whether the worker had any power to lower it, what evidence had justified the threshold, how many cases one employee was expected to supervise, whether rejecting the agent affected performance metrics, and whether the system displayed the evidence needed to verify its conclusion or merely presented a polished summary. The last question proved particularly important. A human can be given final authority in formal terms while being deprived of the epistemic materials required to exercise it.

The vendor's responsibility was more difficult to specify because the hypothetical error might have arisen from model behaviour, interface design, inadequate guardrails, or the employer's own configuration. Several technologists resisted a diffuse theory in which responsibility simply expanded until every participant in the technical chain became responsible for every failure. Others replied that the alternative was often a highly convenient concentration of blame on the one actor with the least control over system selection and workplace design: the frontline employee whose approval remained visible in the audit trail.

The session concluded without a formula for liability and did not attempt one. It produced instead a principle about moral and organizational responsibility: the significance of human approval depends on the conditions under which approval is given. Where the organization controls the system, the workload, the information shown to the reviewer, and the consequences of slowing the process, a final human click cannot fairly be treated as an exculpatory device for everyone upstream. Responsibility may be shared, but it should track meaningful authority rather than the happenstance of who touched the workflow last.

Session Three. The autonomy ratchet: when an agent earns more room to act

The third session turned from failure to success, because several participants argued that successful systems create the more difficult governance problem. The workshop used Microsoft Research's Magentic-UI as its technical reference point. By July 2025 the research prototype allowed users to collaborate on a plan, intervene during execution, require approval for consequential actions, determine how frequently approval should be requested, verify results, and save plans for later reuse. These features made it unusually useful for examining a question that had begun to surface across enterprise AI: not whether an agent should be autonomous, but how a relationship moves from supervision toward autonomy over time.

Participants were given the fictional case of a customer-operations manager who had used an agent for one month to process routine refund requests. During the first week the agent could gather the customer record, check eligibility, and prepare the refund, but it had to ask before taking any action. During the second week the manager allowed it to complete the internal paperwork before seeking approval. By the end of the month the system had processed two hundred cases without an observed error, and the manager entered a standing instruction: for the next month, you may handle refunds of fifty dollars or less without asking me; anything above that amount must still come back for approval.

The room was asked to decide whether anything objectionable had occurred. At first the answer was largely no. The instruction was bounded, intelligible, reversible, and tied to a low-value action the worker understood well. It looked, to many participants, like exactly the kind of calibrated delegation that a mature human-agent workplace should permit. The difficulty emerged when the facilitators altered one fact at a time. What if the system had been correct in two hundred cases but had never encountered a fraudulent claim? What if the company expanded the product line midway through the month? What if a software update changed the model? What if fifty dollars were trivial to the company but significant to a vulnerable customer? What if the worker who set the threshold went on leave and another employee inherited the workflow without having observed the system's earlier performance?

The exercise exposed how easily confidence in a system can become proleptic, treating past performance as though it had already demonstrated reliability under circumstances the agent has not yet encountered. Technologists argued that every supervisory relationship involves some form of generalization from experience and that refusing to expand autonomy until every possible condition has been observed would make delegation practically useless. Human factors researchers replied that the point was not to forbid inference but to define the evidence that justifies it, particularly when the person expected to supervise the system may be learning to rely on it at the same time that the system is receiving more authority.

Participants called this dynamic the autonomy ratchet. Each successful case supplies a reason to reduce friction. Reduced friction increases throughput. Higher throughput makes a return to detailed review more expensive. The system consequently acquires not only more permission but more organizational indispensability, and the employee's ability to move the setting back can diminish even if the interface still contains the button. Several worker representatives noted that a formally reversible delegation may become practically irreversible once staffing levels and performance targets have been redesigned around it.

The session therefore rejected user comfort as a sufficient measure for increasing agent autonomy. Familiarity matters, but it can reflect either well-calibrated trust or habituation. Participants proposed that any increase should consider at least five factors: observed reliability in the relevant class of task, the reversibility of the action, the magnitude and distribution of potential harm, the worker's continuing ability to perform or understand the underlying task, and the practical ease with which autonomy can be reduced again. No agreement was reached on how those factors should be weighted. The important shift was conceptual: autonomy was treated not as a property the machine possesses, but as authority that people and institutions allocate.

Session Four. When oversight becomes theatre: the OpenAI confirmation problem

The fourth session examined a safeguard that had survived every previous discussion: consequential actions should require a person to confirm them. The workshop used ChatGPT agent as the concrete 2025 case because its July launch made the boundary between research and action visible to ordinary users. The system could browse websites, work with connected information, fill forms, edit spreadsheets, and carry out multi-step online tasks, while OpenAI described explicit user confirmation before actions with real-world consequences and active supervision for certain sensitive tasks, including some email activity. Participants were asked to consider whether these controls remain meaningful when they move from occasional use into routine work.

The fictional employee was a partnership operations manager responsible for vendor onboarding across several markets. The agent researched each vendor, assembled the required public information, prepared internal spreadsheets, completed draft forms, and generated routine follow-

up emails. During the first week the employee watched every sequence closely, read the messages, checked the forms against source materials, and frequently interrupted to make small corrections. By the fourth week the system had become familiar, the obvious errors had become rare, and the employee's organization had begun to assume that the same person could process roughly three times the previous volume. The confirmation prompts remained. The character of the confirmation did not.

Participants were shown a screen sequence in which the employee approved twelve routine actions in seven minutes. Nothing in the interface had changed from the first week, and the human was still formally in control. Yet the worker no longer reconstructed each decision from the underlying information. The approval had become closer to a presumption that the system was correct unless something conspicuous looked wrong. Several participants immediately recognized the pattern from aviation, medicine, cybersecurity, and other domains in which reliable automation changes the vigilance of the person asked to supervise it.

The session then posed its most provocative question: when does mandatory approval become less honest than admitting that the automated system is making the routine decision? A number of participants resisted the premise. Even imperfect human review may catch rare but consequential mistakes, and the fact that people sometimes approve reflexively is not an argument for removing their veto. Others replied that the danger lies not in the veto itself but in the institutional use made of it. A company may rely on human approval to characterize a process as human-led while simultaneously setting workloads that make independent verification impossible. The safeguard then becomes less a form of control than a liability architecture.

The Forum called this oversight theatre: a situation in which the human remains conspicuous enough to legitimate the process but under conditions too attenuated to govern it meaningfully. Participants stressed that the term should be used carefully. Not every quick approval is theatrical, and expertise often allows experienced workers to verify routine matters rapidly. The relevant question is whether the organization has preserved the material conditions of judgment: access to evidence, adequate time, a manageable span of supervision, permission to disagree, and a genuine ability to stop or alter what will happen next.

This produced one of the few propositions that attracted almost no resistance. Oversight is work. If a company relies on a human reviewer as the reason an agentic process is safe, responsible, or compliant, then the time and expertise required to review well cannot be treated as an invisible remainder after productivity gains have been calculated. The organization cannot count the human twice, first as the safeguard that justifies autonomy and then as the efficiency opportunity from whom all of the review time may be removed.

Session Five. What happens to expertise? GitHub Copilot and the apprenticeship problem

The fifth session moved away from individual errors and toward a slower institutional risk. GitHub's Copilot coding agent had entered public preview in May 2025 with a deliberately familiar metaphor: assign an issue to Copilot as one might assign it to another developer, allow the agent to work in the background, and review the pull request when it returns. GitHub emphasized low-to-medium complexity work in well-tested codebases, including bug fixes, refactoring, tests, and documentation, while preserving human review and existing repository protections. The product was therefore an unusually clear example of the promise attached to agentic work: machines perform the repetitive or lower-value tasks, leaving people free for more creative, complex, and consequential work.

Participants were asked to imagine a twenty-four-year-old software engineer in the first year of professional employment. Before the coding agent was introduced, the engineer spent a large

portion of the week tracing bugs, extending tests, refactoring ungainly code, updating documentation, and making small feature changes under review from more experienced colleagues. Six months later many of those issues were routinely assigned to Copilot. The junior engineer now wrote clearer issue descriptions, monitored several agent tasks, reviewed completed pull requests, requested corrections, and spent more time on design meetings and higher-level feature work. By conventional productivity measures the new arrangement was successful.

The facilitators then asked what the junior engineer had stopped doing. The answer was uncomfortable because many of the removed tasks were precisely the activities through which software engineers acquire tacit knowledge of a codebase: encountering strange edge cases, discovering why an apparently foolish pattern exists, learning to recognize brittle abstractions, following a failure through several layers of a system, and developing the intuition that allows an experienced reviewer to notice that something technically plausible is wrong. None of these activities is glamorous. Their pedagogical value has historically been hidden inside the ordinary mess of junior work.

The same problem was then translated into other professions. A junior lawyer who no longer performs first-pass research may gain time for client work but lose repeated exposure to the sources from which legal judgment is formed. A translator who increasingly reviews completed machine translations may become an excellent editor while encountering fewer moments in which meaning must be constructed from first principles. An analyst who supervises agent-built models may learn to interrogate outputs without acquiring the same tactile familiarity with the assumptions buried in the model. Participants resisted any romantic defense of drudgery, but they also rejected the idea that every low-status task is developmentally empty.

The discussion became particularly pointed when senior professionals in the room were asked to identify which parts of their own expertise had been acquired through work they would now be tempted to delegate. The list was long. Several participants described the risk as a form of institutional deskilling, not because every worker would become less capable, but because organizations might erode the apprenticeship substrate on which later expertise depends while enjoying the short-term appearance of a more senior workforce.

No participant proposed preserving inefficient work merely because previous generations had endured it. The challenge was more exacting: if a task has both productive and developmental value, an organization that automates the productive component should identify how the developmental component will be replaced. The Forum called this the expertise pipeline problem. It may become one of the most consequential dignity questions in agentic work because professional standing depends not only on having a human title but on retaining the human capacity to know what one is doing.

Session Six. Designing a dignified human-agent workplace

The final working session required participants to stop criticizing existing arrangements and design one. Each table received the same fictional brief: a mid-sized insurer intended to introduce an agentic workflow in 2026 for routine policy servicing, using capabilities no more advanced than those already publicly demonstrated in 2025. The agent would be able to gather information across internal systems, prepare correspondence, update records, and complete defined low-risk actions. Human staff would handle exceptions, consequential changes, and disputes. The board wanted substantial productivity gains but had also committed to a workplace model in which employees remained genuinely responsible professionals rather than nominal overseers of an automated process.

Teams were given a blank one-page instrument called the Human-Agent Work Mandate. They had to specify the objective of the workflow, the actions the agent could perform without approval, the actions it could prepare but not execute, the events that required escalation, the evidence a worker must be shown before approving a consequential action, the maximum volume of agent work one person could supervise, the circumstances under which a worker could reduce or suspend autonomy, the records available after an incident, and the party responsible for reviewing whether the arrangement was eroding human expertise over time. The exercise was intentionally bureaucratic. Participants found that many dignity questions become much easier to evade when left at the level of aspiration and much harder when someone must write the rule that governs Tuesday afternoon.

One table gave workers broad discretion to reduce the agent's autonomy whenever they believed the context had changed. Another objected that a large organization cannot operate reliably if every employee maintains a different private threshold for machine action. A third proposed team-level autonomy bands with a protected individual right to stop a specific sequence. The employment lawyers asked whether exercising that right could affect performance evaluation. The technologists asked whether the system would retain enough state to stop safely midway through a multi-step task. The human factors researchers asked how many alerts would appear before the right to intervene became another unused feature.

The most difficult field on the mandate was labelled Human Purpose. Each group had to state, in one paragraph, what distinct contribution the human was expected to make and why the organization should preserve it. Answers such as accountability, oversight, and judgment were rejected unless they were made operational. If the human's purpose was to catch errors, what information would allow them to do so? If the purpose was ethical judgment, which decisions would remain genuinely open to judgment? If the purpose was empathy in difficult customer cases, could the worker depart from the recommendation without being penalized? If the purpose was professional responsibility, what authority accompanied that responsibility?

By the end of the session, the phrase human in the loop had almost disappeared from the room's vocabulary. Participants spoke instead about where judgment sat, which authority had been delegated, what evidence was available, how reversibility worked, and whether the employee could still exercise the expertise for which the institution claimed to value them. The Forum regarded this linguistic shift as more than cosmetic. A vague safeguard had been replaced by a set of questions that could be answered badly, argued over, audited, and improved.



The human role should be stated with enough precision that an employee can tell when the organization has quietly removed it.

PART THREE

III. The dignity problem

The final afternoon returned to the Institute's reason for convening the discussion at all. Productivity, safety, and legal responsibility could each justify serious attention to human-agent work, but participants were asked what changes when the problem is framed through dignity. The answer was not that dignity supplies a sentimental objection to automation. It supplied a more exact account of what can be lost while employment formally remains intact.

Judgment and epistemic standing

In many occupations, a person's standing at work is tied to the fact that their judgment counts. They are expected to notice, interpret, weigh, and decide rather than merely transmit a conclusion produced elsewhere. Agentic systems can strengthen that capacity by removing clerical burden and presenting information more effectively. They can also hollow it out if the person is left with formal approval but little independent epistemic purchase on the decision. Several participants used the phrase epistemic standing to describe the difference: the worker is not simply present, but recognized as someone whose reasons and disagreement have institutional force.

Authorship

The Forum distinguished responsibility for supervising a system from authorship of every act performed under that supervision. Human work has never been individually pure; it is produced through teams, precedents, templates, software, organizational rules, and hierarchy. What agentic systems add is the possibility that much of the cognitive production of a decision may occur outside the person while the final output continues to be attributed to them as though it expressed their own professional judgment. Participants regarded accurate attribution of human and machine contribution as a dignity interest because a worker should neither have their own contribution invisibly appropriated nor be made the fictitious author of a conclusion they did not meaningfully form.

Capability and the right to remain competent

The expertise discussion introduced a temporal dimension that ordinary governance frameworks often miss. Dignity at work is not exhausted by the worker's experience in the present moment. It includes whether the organization preserves a plausible path by which people can acquire, maintain, and exercise competence. A workplace can remain pleasant, well paid, and apparently human-centered while steadily reducing the opportunities through which human judgment is learned. The Forum stopped short of describing this as a right to perform inefficient tasks, but several participants argued that organizations introducing agentic systems should be able to explain how the capabilities they continue to demand from people will actually be cultivated.

Voice and organizational power

The most persistent correction made by worker representatives was directed at the language of partnership between a person and an AI. In employment settings the individual rarely determines the relationship alone. The employer selects the technology, decides which data and tools it can access, sets productivity expectations, establishes escalation rules, and evaluates the employee afterwards. A genuinely dignified human-agent workplace therefore requires worker voice not only at the interface but in the design of the work itself, particularly where the employee will later be expected to answer for what the system does.

PART FOUR

IV. Where the Forum divided

1. Whether autonomy should be earned. Several participants favored progressive delegation in which an agent receives wider authority after reliable performance under supervision. Others argued that capability and authority should remain conceptually separate, since a system may

become technically excellent at a task that an organization, profession, or society nevertheless decides should remain subject to human judgment. The disagreement was not resolved, but it clarified that evidence of competence is necessary for autonomy without being sufficient to justify it.

2. Whether human approval should ever be mandatory. One group defended contemporaneous approval for consequential and irreversible decisions as a minimum safeguard. Another argued that mandatory approval becomes morally ambiguous when organizations increase throughput beyond the point of meaningful review while preserving the human signature as a shield. Both sides agreed that the existence of approval should never be treated as self-proving evidence of independent human judgment.

3. Whether workers should control autonomy settings. Worker representatives argued that an employee expected to answer for an agent should have a corresponding right to reduce its autonomy. Employers and technologists raised concerns about consistency, operational reliability, and the possibility that individualized settings would make systems difficult to govern. A narrower proposition attracted more support: workers should at minimum possess a protected capacity to stop or escalate a specific sequence without being penalized for doing so in good faith.

4. Whether preserving expertise requires preserving inefficient work. Some participants regarded the expertise pipeline argument as vulnerable to nostalgia, noting that organizations have always removed obsolete tasks when better tools became available. Others replied that apprenticeship is costly by definition and that a firm which removes the activities through which judgment develops must replace their developmental function rather than assume expertise will continue to appear spontaneously. The Forum recorded the disagreement as a research question rather than a principle.

5. Whether transparency about machine authorship will scale. Participants supported accurate attribution in principle but disagreed about how granular it should become in ordinary work. One group favored visible records showing which portions of a consequential decision were generated, recommended, or executed by an agent. Others cautioned that excessive attribution could become another compliance ritual whose informational value collapses under volume. The unresolved question was what a person affected by a decision, and what a worker asked to defend it, genuinely needs to know.

PART FIVE

V. The Maroma Principles for Dignified Human-Agent Work

Adopted at the closing session of the DDI Fall Forum, 24 October 2025. The propositions are intended as conditions for meaningful human participation in agentic work rather than as a general code for artificial intelligence.

1. Human presence is not human control. The existence of a person somewhere in a workflow should not, by itself, be treated as evidence of meaningful human oversight. Control requires sufficient information, competence, time, discretion, and authority to alter the outcome rather than merely acknowledge it.

2. Responsibility should follow meaningful authority. A worker should not bear responsibility for an agentic action merely because they occupied the final approval position in a process whose

material decisions, constraints, or evidence they could not reasonably examine or change. Organizational responsibility must remain visible where the organization determined the conditions under which the worker acted.

3. Oversight is work. Where an organization relies upon human supervision to make an agentic system safe, reliable, accountable, or compliant, the time, training, information, and manageable workload required for effective supervision should be treated as substantive work rather than as friction to be engineered away.

4. Delegation should be granular. Authority to research, recommend, prepare, communicate, execute, and commit are distinct forms of authority. Permission at one level should not silently become permission at another merely because the system is technically capable of continuing the sequence.

5. Agent autonomy should be legible and reversible. Workers should be able to understand which actions an agent may take without intervention, what triggers escalation, and where approval remains required. Increases in autonomy should be deliberate, and any person expected to answer for the system should have a practical route to reduce, suspend, or challenge that autonomy when conditions change.

6. Human approval should not create fictitious authorship. Organizations should not attribute a substantially machine-produced judgment to a human solely because the human confirmed, transmitted, or failed to interrupt it. Approval may establish responsibility for supervision without establishing that the underlying reasoning was authored by the person.

7. Human judgment requires the conditions for judgment. A nominal veto is not meaningful where a person lacks the evidence, time, expertise, independence, or organizational permission required to use it. The more consequential the decision, the stronger those conditions should be.

8. Workers should have voice in agentic work design. People expected to work with, supervise, correct, or answer for artificial agents should participate in decisions concerning the agent's authority, escalation rules, monitoring, performance expectations, and allocation of responsibility, particularly where those decisions materially alter the character of their professional role.

9. Human capability is an organizational asset. Deployment decisions should consider not only immediate productivity but whether the resulting distribution of work preserves the acquisition, maintenance, and exercise of human expertise. When a developmental task is automated, the organization should identify how its developmental function will be replaced.

10. Human involvement should have a stated purpose. A person should not be retained in a workflow principally to legitimate an automated process or absorb responsibility for it. Where human involvement is required, the organization should be able to state what distinct human judgment, authority, relationship, or value that involvement is intended to provide, and structure the work so that the purpose can actually be fulfilled.

PART SIX

VI. Research agenda

1. What counts as meaningful human oversight? A measurable taxonomy is needed that distinguishes presence, review, verification, control, judgment, and authority in real workplace

settings. The Forum regarded this as foundational because regulation, procurement, and organizational policy increasingly rely on human-in-the-loop language without a stable account of what the human is actually doing.

2. How should autonomy be calibrated over time? Empirical work should test whether autonomy can be progressively expanded using task risk, reversibility, observed reliability, worker expertise, and consequence, and should examine what evidence is sufficient to move from approval before action to audit after action.

3. When does supervision stop working? Human factors research should measure the relationship among system reliability, review volume, time pressure, automation bias, habituation, and the probability that a worker will detect a rare error. The Forum considered this one of the most immediately answerable questions with direct consequences for workplace design.

4. How many agents or agent actions can one person meaningfully supervise? Enterprise visions increasingly assume that one employee may oversee multiple autonomous processes at once. There is little basis for determining a defensible span of supervision, and the answer is likely to differ by task complexity, consequence, interface design, and the availability of evidence.

5. Who should control the autonomy setting? Comparative workplace research should examine systems in which employees, managers, administrators, vendors, or the agent itself determine when approval is required. Particular attention should be paid to whether the person who bears responsibility can meaningfully reduce autonomy and what happens when exercising that power conflicts with productivity expectations.

6. How should responsibility be allocated across the workplace system? A practical framework is needed that separates system selection, configuration, task delegation, supervision, individual approval, organizational incentives, and technical failure rather than assigning responsibility to whichever human appears last in the event log.

7. What happens to expertise under sustained agent supervision? Longitudinal studies should examine whether workers who increasingly review agent-produced work develop, maintain, or lose the capacity to perform and judge that work independently, and whether the effect differs between experienced professionals and people still acquiring expertise.

8. What does agentic apprenticeship look like? If traditional junior work increasingly shifts to artificial agents, organizations will need alternative mechanisms through which workers acquire tacit knowledge, pattern recognition, judgment, and professional confidence. The Forum regarded this as a central future-of-work question that should be studied before the expertise pipeline is visibly damaged.

9. What rights should accompany agent supervision? Workplace policy and employment law should examine whether workers expected to answer for artificial agents require corresponding rights to inspect, interrupt, override, reduce autonomy, obtain records, report system failures, and refuse an agent-generated action in good faith without retaliation.

10. Can machine contribution be attributed without producing disclosure fatigue? Research is needed into how organizations can record and communicate the respective roles of human and machine in consequential work without turning attribution into a barrage of notices that no one reads. The goal is reconstructability and accurate authorship, not decorative transparency.

ANNEX A

Composition of the Forum

Thirty participants attended by invitation under the Chatham House Rule. In accordance with that rule, roles are described and names are withheld. The composition was intentionally designed to create disagreement across technical, philosophical, organizational, and worker perspectives rather than to assemble a room of professional AI-governance specialists who already shared the same vocabulary.

- Four scholars working on moral agency, responsibility, delegated authority, and human-machine interaction.
- Three researchers in human factors, automation bias, trust calibration, and supervisory control.
- Three specialists in labour and employment law, including one with expertise in professional responsibility.
- Two researchers working on social dialogue, worker voice, and technological change.
- Four technologists working on agent orchestration, enterprise agents, human-in-the-loop systems, and interaction design, attending in a personal capacity.
- Three organizational leaders responsible for implementing AI systems within workplaces.
- Two specialists in professional ethics and regulated occupations.
- Two organizational psychologists studying expertise, judgment, and skill formation.
- One specialist in occupational safety and systems accountability.
- Six workers or frontline professionals whose roles involved substantial AI-assisted or automated workflows, drawn from software development, translation and localization, customer operations, professional services, human resources, and financial or claims processing.

The Forum regarded the final category as indispensable. Its object was not to determine how institutions could keep humans in the loop. It was to examine what happens to the human once they are placed there, and the discussion repeatedly changed when workers described the distance between a control that exists in software and a control that can actually be exercised in an employment setting.

ANNEX B

Note on method

The Forum sat for three days in six working sessions with a closing synthesis at which the Maroma Principles were adopted. Sessions were unrecorded. The rapporteur took contemporaneous notes and circulated a draft for correction of fact rather than of interpretation. Participants were not asked whether agentic AI would transform work, a question the Forum regarded as too broad to discipline serious discussion. Each session began instead with a specific division of labour among a worker, an artificial agent, and an employing organization.

The company scenarios were constructed from product capabilities publicly described; they were deliberately plausible rather than predictive. The Salesforce exercise used Agentforce 2dx as a reference point for proactive, background enterprise agents. The DeepL exercise used the controls DeepL described for monitoring, stopping, inspecting, validating, and approving agent

activity. The Microsoft exercise used Magentic-UI's co-planning, co-tasking, action approvals, configurable approval frequency, verification, and saved-plan capabilities. The OpenAI exercise used ChatGPT agent's combination of research and action with explicit confirmation for consequential actions and active supervision for some sensitive activities. The GitHub exercise used Copilot coding agent's ability to accept an issue, work asynchronously, and return a pull request for human review.

For each scenario, participants were required to identify separately the objective, the judgment, the action, the authority to intervene, the information available to the human, the organizational incentives shaping the human's behavior, and the location of responsibility after failure. The method produced repeated disagreement about whether the worker in a given case remained meaningfully in control. That disagreement was not treated as a weakness of the exercise. It became its principal finding.

The phrase human in the loop describes a position in a system. It does not tell us whether the person in that position can understand what is happening, whether they have time to judge it, whether they are allowed to disagree, whether their expertise is being maintained, or whether the institution will listen when they intervene. The Forum concluded that workplace AI governance will require considerably greater precision if the human is to remain more than a ceremonial remainder in an increasingly automated process.

ANNEX C

Documentary basis for the scenarios

The following public materials supplied the technological boundary conditions for the named workshop cases.

Salesforce, “Salesforce Launches Agentforce 2dx”. 6 March 2025. Salesforce. Described proactive Agentforce capabilities, data-triggered action, autonomous background operation in business processes, and tools for configuring, testing, monitoring, and optimizing agent performance.

DeepL, “Agentic AI for business”. September 2025 product status. DeepL AI Labs. Described DeepL Agent as in beta with selected customers and identified multi-layered controls that allowed users and administrators to monitor and manage agent tasks, stop and inspect the agent, and require human validation and approvals.

Microsoft Research, “Magentic-UI: Towards Human-in-the-loop Agentic Systems”. July 2025. Microsoft Research Technical Report MSR-TR-2025-40. Described co-planning, co-tasking, action approval, answer verification, memory, and multi-tasking as interaction mechanisms for human involvement in agentic systems.

OpenAI, “Introducing ChatGPT agent: bridging research and action”. 17 July 2025. OpenAI. Described a system combining research and online action, with explicit user confirmation for consequential actions, the ability to intervene, and active supervision for certain critical tasks.

GitHub, “GitHub Copilot coding agent in public preview”. 19 May 2025. GitHub. Described assigning issues to Copilot, autonomous background work in a development environment, creation of pull requests, testing and linting, and subsequent human review.