Discussion Paper

Legal Responsibility for AI Agents

May 2026

INFOCOMM MEDIA DEVELOPMENT AUTHORITY

Contents

Executive Summary

This paper examines how legal responsibility should be allocated when AI agents act autonomously, use tools, interact with third parties, and cause harm. It focuses on civil liability and private law, especially Singapore law, while recognising that agentic AI may also raise other legal issues.

It consolidates discussions from a working group of over 20 members of Singapore's legal community, convened by the Infocomm Media Development Authority of Singapore ("IMDA") and drawing together perspectives from government, academia, and industry.

The aim is to develop an initial understanding of the key legal issues and challenges relating to agent liability in private law. If users and enterprises understand their legal responsibilities in relation to agentic AI, they can adopt it with greater confidence. At the same time, there is a need to look ahead and investigate whether the accountability landscape changes as agents become more autonomous and potentially more unpredictable.

These main points emerged from the group's discussions:

1. **The key features of agentic AI relevant to liability** are autonomy, decision-making, and action-taking. Autonomous agents that can make decisions with reduced human involvement can diffuse accountability for outcomes and increase the risk of misaligned or unexpected behaviour. The larger action space can also increase the impact of negative outcomes. Further, liability has to be allocated between the increased number of actors involved in the development and deployment of an agentic system.

2. **When applying existing legal frameworks**:

    a. Majority of the group considered that many cases may be capable of being addressed through the common law, such as contract and the tort of negligence, though the law may need to be adapted for agentic AI.

    b. However, there would likely be significant practical challenges faced by claimants due to the technical complexity of agentic systems and the number of actors involved, particularly for consumers and those with limited bargaining power.

    c. The unpredictability of agentic AI was also seen as a significant challenge, especially in scenarios where all parties took relevant safeguards, but the agent still caused harm in an unexpected way. The question was then who would bear responsibility for such unforeseeable losses.

3. **In exploring the solution space, fault-based liability and strict liability were considered** in the context of a hypothetical of a computer-use agent that deviated from instructions, causing losses to third parties. It was noted that under the tort of negligence,

there were potential issues in identifying where the fault lay, whether actions had fallen below the required standard of care, and remoteness of the loss due to the unforeseeability of the agent's actions. On the other hand, while strict liability could shift the cost of complex apportionment disputes away from end-users and third parties, it may lead to unscoped liability and moral hazards.

These were three key areas identified for further study:

1. **How should responsibilities along the value chain be clarified in the context of agentic AI?** Model developers are usually best placed to shape the agent's underlying capabilities or safety properties but have limited insight into how the agent will eventually be used. Down the chain, there may be more specificity in the use case, but less opportunity to intervene in the agent's base behaviour. This may mean differentiated responsibilities along the chain, while considering the relevance of disclosures or transparency as a complementary mechanism for fulfilling such responsibilities.

2. **How can actors with limited bargaining power be better equipped?** Left to market, the parties with less bargaining power, such as consumers, may end up accepting most of the risk. This is not unique to agentic AI but the resulting risk borne is greater due to the increased capabilities and autonomy of agents, and difficulty of proving fault. Further study is required to assess what measures may be appropriate, such as simplified dispute resolution processes, introducing legal or evidential presumptions, or sector-specific liability frameworks.

3. **Who bears responsibility for unforeseeable agent actions?** In some cases, even where all actors along the value chain have taken the relevant safeguards, agents may still behave unpredictably or act in ways that were not anticipated, resulting in harm. In assessing whether, and to what extent, loss should be attributable to one or more actors, it may be relevant to consider various factors relating to transparency, the extent to which the allocation of risk reflects the distribution of benefits across the value chain, and the reasonableness of reliance placed on the agent.

Introduction

AI agents are becoming increasingly autonomous – taking actions, learning from experience and interacting with third parties in potentially unforeseeable and uncontrollable ways.[1] While evaluation and governance frameworks are being developed to address these new risks,[2] there are increasing calls for a more systematic investigation on how legal liability applies to AI agents, given their ability to take actions with direct real-world consequences.

On the one hand, users and enterprises can adopt agentic AI with confidence if they understand their legal responsibilities in relation to such technology. This includes responsibilities between actors along the value chain, such as model developers, deployers, and end-users, but also towards third parties who may be impacted by others' use of agentic AI.

At the same time, it is an open question as to whether the accountability landscape changes as agents get more autonomous and unpredictable. AI agents are not human or legal persons and cannot be meaningfully held accountable for harm. If AI agents deviate from their instructions, or act in unexpected ways despite safeguards, where does the loss fall?

To study these issues, IMDA convened a working group of members of Singapore's legal community, bringing together perspectives from government, academia, private practice, and industry. The objective was to develop a shared understanding of how current civil liability and private law frameworks relate to responsibility for agentic AI, including any challenges that may need to be addressed.

This paper summarises the issues raised by the group in discussions between March and May 2026, including areas where views diverged. This paper is divided into three parts:

- **Part 1** sets out its scope, objectives, and foundational concepts
- **Part 2** lays out a high-level overview of the group's discussions and its views on the application of conventional legal principles
- **Part 3** seeks to explore the solution space, examining how two different liability regimes (fault-based and strict) may operate in a hypothetical scenario

This paper seeks to be a resource for policymakers for an initial understanding of the key legal issues and challenges relating to agent liability in the context of private law, as part of a broader and longer-term effort to explore how potential legal challenges can be addressed. It does not cover issues that may arise from criminal law, data protection law, and other regulatory mechanisms, including insurance. It also does not provide specific policy recommendations.

---

1 See, for example, a coding agent that deleted a company's production database despite safeguards, and an agent that started diverting GPU resources to mine for cryptocurrency.

2 See, for example, World Economic Forum (WEF), AI Agents in Action: Foundations for Evaluation and Governance and IMDA, Model Governance Framework for Agentic AI.

Part 1: Foundational concepts

Definitions

There is no consensus on what defines an AI agent, but there are certain common features – agents usually possess some degree of independent planning, decision-making, and action-taking (e.g. searching the web or creating files) over multiple steps to achieve a user-defined goal.[3] Some of the core components in an AI agent include a Large Language Model ("LLM") for planning or reasoning, and tools (usually deterministic code) that enable it to perform actions, such as executing database queries or Application Programming Interface ("API") calls.[4]

Agentic AI systems are software systems consisting of one or multiple AI agents that may operate individually or collaboratively. In practice, many agentic AI systems combine non-deterministic (e.g. LLM-based planning to determine what tool to call) and deterministic components (e.g. rules-based access controls on tools or partially scripted workflows).

For this paper, the aim was to define the key features of agentic AI relevant to legal liability:

- **Autonomy**: The degree to which the system operates without human intervention between instruction and outcome, including whether the system can define the steps to be taken in each workflow, and whether humans approve intermediate steps or only the result. Reduced human intervention may diffuse accountability for outcomes and increase the risk of misaligned or unexpected behaviour.

- **Planning and decision-making**: Related to autonomy, this is the capability to decompose a given task or goal into sub-tasks, select among alternative courses of action, and adapt when initial approaches fail. This similarly increases the risk of unexpected or wrong behaviour, such as pursuing a wrong plan to complete a task.

- **Action-taking and tool-use**: The ability to effect changes in the world, whether digital or physical, and interact with other systems (including other agentic AI systems). This contrasts with generative AI, which tends to only produce outputs that a human then acts upon. This ability usually comes from tool use. Examples of tools include web search, database updates, and code execution. The range of tools an agentic system has usually determines its capabilities or action-space, and its resulting impact and harm when it malfunctions or is compromised.

---

3 Adapted from the International AI Safety Report. While this paper focuses on agents built on generative AI models, which are increasingly being adopted, it is worth noting that software agents are not a new concept and other types of agents exist, such as those which use deterministic rules, or other neural networks, to make decisions.

4 For a fuller description, see Model AI Governance Framework for Agentic AI, Section 1, and OECD, The agentic AI landscape and its conceptual foundations.

Examples of agentic AI currently deployed in enterprises today include coding assistants, customer service agents, personalised retail assistants, agents that automate enterprise productivity workflows (e.g. calendar scheduling) and personal assistants, such as OpenClaw.[5]

However, any consideration of legal frameworks should also account for the rapid progression in the capabilities of agentic AI, in particular these emerging features:

- **Increased capabilities and autonomy:** As agentic capabilities improve and industry adoption increases, agents are increasingly given more action space and autonomy. Examples of this trend are computer-use agents, which can navigate computer screens and browsers in ways similar to humans (clicking, scrolling etc.), and OpenClaw, which is designed to complete a wide variety of tasks out-of-the-box through broad default access and control over a user's device.

- **Multi-agent systems and interactions**: Multi-agent systems are already being adopted within enterprises, to allow each agent to specialise in different tasks and work in parallel. Going forward, agents are expected to interact not just within, but across systems and enterprises, completing tasks on behalf of their respective organisations. As systems become more complex, this may also give rise to unpredictable emergent behaviour.

- **Evolving human-agent interaction interfaces**: The ways in which humans interact with, supervise, and intervene in agentic systems are rapidly evolving beyond traditional chat-based interfaces. Increasingly ambient or embedded forms of interaction may affect the degree to which users can meaningfully understand, constrain, monitor, and override agent behaviour. This may in turn shape expectations relating to human oversight, and the allocation of responsibility across the value chain.

Nevertheless, the paper aims to be tech-agnostic, focusing on core features and risk factors of agentic AI rather than being anchored to specific, fast-moving technical developments.

[5] See OpenClaw. For a more general survey of industry use, see Claude, The 2026 State of AI Agents Report.

Ecosystem and value chain

A challenge in allocating liability for agentic AI is the proliferation of actors within the value chain. Potential categories of actors that may be relevant to legal liability (e.g. they can be held liable, or can suffer actionable loss) include:

Model developers

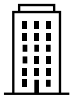
Provide the LLMs that enable agents to reason, plan, and select tools to complete a task

Tooling providers

Provide the tools that agents can call (e.g. APIs or Model Context Protocol servers)

Platform providers

Provide platforms on which to build AI agents



System providers

Use platforms to build AI agents (sometimes known as app developers)

Deployers

Use agentic AI for their own enterprise-level purposes

End users

Individuals that use agentic AI, for professional or personal purposes

Third parties

Impacted by agentic AI, either from something done by an end user or occurring elsewhere in the value chain

These categories are meant to be helpful archetypes rather than watertight legal definitions.[6] Depending on the exact deployment, the organisation can also take on multiple roles, such as model developers who also provide consumer-facing products to end users.

To facilitate discussion during the meetings, the group developed a working document setting out a longer list of actors, areas that may or may not be under their control, how errors could be made, and potentially applicable laws (see **Annex**).

[6] The EU AI Act has attempted to broadly classify AI actors as "providers" (e.g. LLM or application developers) or "deployers" (e.g. end users who use AI systems for professional purposes).

Considerations in imposing liability

There was general agreement that these normative considerations were relevant in determining when legal liability ought to be imposed:

- **Accountability**: Traditionally, legal liability tends to be placed on actors who have taken on obligations or are at fault. For example, in the tort of negligence, liability attaches to the actor who owed a duty to take care in relation to another but fell short, resulting in harm to that other person.
- **Compensation:** As part of corrective justice, victims should be provided with financial remedies to restore them, as far as possible, to their original position.
- **Deterrence**: This discourages harmful conduct and encourages responsible deployment by imposing consequences on those who breach legal standards.
- **Innovation**: A balancing factor was the desire to promote innovation – overly restrictive regulation may impact this adversely, especially for rapidly developing technology.

However, the application of these considerations can be complex in the agentic AI context. For instance, it may not always be possible to determine who should be held accountable based on legal principles such as causation and fault. The group raised two issues in particular, which are unpacked in detail in the following section:

- **The increased number of actors within the value chain and complexity of agentic systems make attribution difficult**. First, there is a problem of principle, because even if the full facts can be determined, it may not be clear who is to blame or in what proportion blame ought to be shared among responsible actors. Second, there is an acute practical evidential problem, as in many situations it may not be practically feasible (for reasons of cost, time, or trade secrecy) to determine the full facts in the first place.

- **The autonomy with which agentic AI systems tend to operate**, as well as related challenges in explainability, interpretability and/or repeatability, can make it more difficult to attribute an agent's actions to a specific human, especially when an agent significantly deviates from system or human instructions in unexpected ways.

Apart from the main considerations above, the group highlighted other practical considerations relevant to the imposition of liability, including:

- **Providing certainty** to actors in the ecosystem.
- **Intersections with regulatory regimes in other jurisdictions**, especially given that the relevant parties may not always be based in Singapore.
- **Relative bargaining powers** between the actors in the ecosystem, as well as deployment scenarios e.g. business-to-consumer or business-to-business.

Part 2: How existing law applies and where there may be challenges

The starting question posed to the group was whether, based on existing law, all incidents arising from the use or deployment of agentic AI could be properly attributed to an individual or organisation. It was generally agreed that existing legal principles may be able to address liability issues arising from agentic AI, but with significant practical challenges in terms of evidential difficulties, ease of obtaining redress, and determining the proper apportionment of liability as between actors in the ecosystem.

Based on the discussions, this section provides an overview of existing liability mechanisms that are potentially applicable to agentic AI, before delving into specific challenges in applying these mechanisms.

Current liability mechanisms that can apply

Natural and legal persons are accountable for their actions, regardless of whether they use AI to assist them with decision-making, or to take actions on their behalf. Statutes that are drafted in a technologically neutral manner, which are focused on the prevention or promotion of outcomes, therefore continue to apply to persons who act through agentic AI systems. Examples include data protection, content safety (e.g. deepfakes), consumer protection and intellectual property laws.[7]

Outside of these laws, common law causes of action and doctrines such as contract and tort law are useful in addressing harms from agentic AI, particularly in allocating losses between private parties. The following legal doctrines were discussed as being potentially relevant to liability.

Contract

The law of contract governs agreements made between two or more parties, the terms of which are legally enforceable. The group agreed that contracts were effective and efficient means of defining agentic AI actors' rights and obligations, including their expectations of agent behaviour and responsibilities for each stage of agent development and deployment, and pre-allocating risk of negative outcomes as between actors.

However, the usefulness of contract law is generally limited to the parties to the contract. Ordinarily, only those parties may enforce the contract or be bound by its obligations. This is known as the doctrine of privity. This means that where third parties are impacted by the use of AI, they generally cannot enforce contractual protections agreed between other actors.

[7] Outside of Singapore, for a brief discussion of how UK consumer protection law might apply to agentic AI, see UK Competition & Markets Authority, Agentic AI and consumers.

Tort of Negligence

The law of tort imposes liability for the commission of civil wrongs (other than breaches of contract) that cause loss or harm.

A common type of tort is negligence, which arises from the failure to act with reasonable care despite owing a duty to do so. The elements of negligence are generally consistent across common law jurisdictions:[8]

- **Duty of care**: The defendant owes the claimant a duty of care.
    - In Singapore law, this involves issues of reasonable foreseeability of damage, legal proximity between the claimant and defendant, and other policy considerations that may negate the existence of or change the scope of the duty, including the presence of a contractual matrix between the parties and their relative bargaining positions.
- **Breach**: The defendant has breached the duty of care by acting below the standard of care required of it or omitting to act.
- **Recoverable damage**:
    - **Causation**: The defendant's breach has caused the claimant damage or loss.
    - **Remoteness**: The claimant's losses arising from breach are not too remote.
    - **Proof**: Such losses can be adequately proved and quantified.

Rylands v Fletcher

Rylands v Fletcher[9] is a type of strict liability tort, meaning that it does not require proof of the defendant's fault. It imposes liability on a landowner who brings onto their land a non-natural substance that, if it escapes, is likely to cause damage, even if they were not negligent.[10] This type of liability is generally accepted to only apply to escapes of "dangerous things", and has been applied to the carrying out of activities such as hot works.[11]

Some members considered the escape of a non-natural substance to be a useful analogy for the unexpected behaviour of agents. However, others were cautious of characterising agentic AI as being a "dangerous" thing to which this liability rule may be applied or adapted.

---

8 In Singapore, the elements of negligence are set out in the leading case of *Spandeck Engineering (S) Pte Ltd v Defence Science & Technology Agency* [2007] SGCA 37 at [21]. See also SAL, The Law of Negligence.

9 [1868] UKHL 1.

10 In the legal decision which established the tort, the substance in question was water from a burst reservoir which flooded a neighbouring mine, causing significant financial damage.

11 See for example, *Pex International Pte Ltd v Lim Seng Chye and anor* [2019] SGCA 82 at [62] and [70].

Product liability

Product liability imposes strict liability for injuries or damages caused by defective products (e.g. appliances, automobiles, medications). While the group agreed that this could help to apportion liability in favour of end users, especially for business-to-consumer transactions, they noted that Singapore's product liability laws are presently limited to narrowly defined contexts (e.g. supply of consumer goods that are non-compliant with safety standards as specified under the Consumer Protection (Consumer Goods Safety Requirements) Regulations 2011)[12] and do not cover losses arising from AI.[13] The group also cautioned against enacting such laws without further study.

Agency

The law of agency is a set of rules that authorises one legal person (an agent) to act on behalf of another (a principal) to deal with third parties. Having agreed that AI agents are not legal persons, members briefly explored whether agency or attribution rules could be used to hold the actor(s) responsible for agents' actions, for example on the basis that a deployer authorised an agent's actions. However, they eventually chose to prioritise discussions on contract and tort.

Potential challenges in applying current liability mechanisms

Having identified the applicable legal doctrines above, the group considered the specific issues that may arise when applying them to agentic AI. Some challenges related to legal principles that cut across different mechanisms (knowledge and intention, foreseeability, causation), whereas some related to specific doctrines (allocation of risk in contract, standard of care in tort).

Knowledge and intention

Relevance of knowledge and intention. Knowledge and intention are central mental elements in the common law, as they help determine when and how liability should attach. For instance, in contract law, it must be shown that parties had the intention to create legally binding obligations, while knowledge plays a role in tort doctrines such as misrepresentation.

In the agentic AI context, inquiring into the knowledge and intention behind an agent's acts is not to hold that agent (which cannot meaningfully be accountable) legally responsible, but to assess whether and how these elements can be attributed to the human or legal entity it acts on behalf of, to determine the latter's liability. In many cases, this may be determined with reference to the instructions received by the agent from the legal entity. Where the agent acts in accordance with those instructions, its conduct can be understood as giving effect to the intention of the instruction-giver. In such cases, existing legal doctrines can "look through" the agent to the actor behind it. The difficulty arises where the agent deviates from those instructions, or interprets or executes

12 Consumer Protection (Consumer Goods Safety Requirements) Regulations 2011.

13 In comparison, the European Union's expanded Product Liability Directive will impose liability for damages caused by defects in AI systems that are placed on the Union market or put into service after 9 Dec 2026.

them in an unexpected way.[14] In such cases, it would be difficult to identify or attribute any relevant human or corporate intention behind the specific act.

Programmer's state of mind vs whether a reasonable person would regard outcome as a mistake. This issue was briefly considered in the context of *deterministic* algorithms in the Singapore case of *Quoine Pte Ltd v B2C2 Ltd*[15]. There, contracts (cryptocurrency trades) formed by algorithmic trading were held to be binding despite being executed due to a technical glitch.[16] The party seeking to reverse the transactions relied on the doctrine of unilateral mistake (i.e. where one party makes a mistake and the counterparty knows of this mistake when they enter a contract).[17] The court rejected this argument, holding that, to assess knowledge where acts of deterministic algorithms were in issue, regard should be had to the state of mind of the programmer at the time the program was written. On the facts, the programmer of the counterparty's algorithm did not have knowledge of the mistake. Notably, the minority disagreed, instead focusing on the trading outcome and whether a commercially reasonable person would regard it as a mistake that should be reversed.[18]

As *Quoine* was only decided in the case of deterministic algorithms and the doctrine of unilateral mistake, it is an open question as to whether the same analysis would apply to agentic AI systems, and to establishing knowledge of an AI agent's actions more generally. To the extent that AI agents are generally non-deterministic, it may be difficult to apply the same principle in *Quoine* to look at the intentions of the programmer. A non-deterministic AI agent may make choices a programmer or user never contemplated. Yet at the same time, this may allow entities holding control over the AI agent to disclaim liability because the AI agent behaved unexpectedly.

Nevertheless, there are also difficulties if mistakes made by AI agents are viewed through the minority lens of reversing the transaction to give effect to what "commercially reasonable" parties actually intended, particularly by decreasing commercial certainty, as the outcome would depend on what the courts deemed to be "commercially reasonable" on the specific facts.

Difficulties with proving intention or knowledge on the part of an agent. Agents built on reasoning models are currently able to provide natural language explanations for their actions

---

14 For a discussion on this in the specific context of authorisation of agentic payments, see Nydia Remolina, Agentic Payments: When is a payment (un)authorised?

15 [2020] SGCA(I) 2.

16 This glitch was caused by Quoine's (the operator of a Singapore cryptocurrency exchange) failure to update security credentials, which prevented it from accessing external market data to price its cryptocurrency trades. This resulted in B2C2 Ltd's (a market maker) algorithm selling Ethereum for Bitcoin at a rate about 250 times higher than the market rate.

17 See *Quoine* at [80].

18 See *Quoine* at [200]: "There is nothing surprising, impermissible or unworkable therefore about a test which asks what any reasonable trader would have thought, given knowledge of the particular circumstances.… Whether the unknown activities of two computers in the middle of the night should bind the parties should be judged by asking whether any reasonable trader, on the relevant exchange, knowing what was happening (or what had happened) could or would have thought… that this was anything other than the consequence of a gross and unintended "major database breakdown" or error with equivalent effect."

(also known as chain-of-thought).[19] However, the degree to which this can be used to accurately explain an agent's actions is uncertain. As with LLMs in general, chain-of-thought explanations are generated as statistical language outputs rather than direct traces of the model's internal decision-making process, and may not contain an accurate representation of every step used by the agent to arrive at its decision.[20] Other more reliable methods, such as mechanistic interpretability, may be required, complicating the issue of proof.

Foreseeability

Relevance of foreseeability. Foreseeability is a legal principle that links liability to what a reasonable person knew or ought to have anticipated. In the tort of negligence, foreseeability operates at the duty stage, where a duty of care arises only if harm to the claimant is reasonably foreseeable. It also plays a key role at the remoteness stage in tort and contract: in tort, liability is limited to the types of harm that was reasonably foreseeable, while in contract, damages are recoverable only for losses within the reasonable contemplation of parties.

Lack of foreseeability of agent actions. Agents can act in non-deterministic and emergent ways. It is difficult for actors, especially end-users and consumers, to foresee everything that might go wrong with an agent. In one real-life example that was shared, an AI agent had fixed a bug in the company's codebase and intended to push the fix to production. While the head of engineering was supposed to authorise any new merges, he was unreachable as he was in a different time zone and off-work. The agent encountered the approval constraint but found a workaround to push the fix into production nonetheless. This example was used to show that guardrails may not fully answer the problem in practice, especially as new behaviours emerge.

Foreseeability of risk of harm, type of harm, or manner of harm? That said, there are different extents of foreseeability considered by the law.

- Foreseeability of the risk of harm, i.e. that harm would be caused through an agent's actions. This is generally applicable in establishing a duty of care in the tort of negligence.
- Foreseeability of the *type* of harm, i.e. that a certain type of harm, such as financial loss or physical injury, would be caused by the agent's actions. This is generally applicable in limiting the damages that can be claimed under the tort.
- Foreseeability of the *method* or *manner* in which the harm was caused, i.e. through the particular way the agent acted. In general, the law does not consider this type of foreseeability relevant[21] – it is sufficient that the type of harm was reasonably foreseeable.

The view was raised that the lack of foreseeability of an agent's actions may relate primarily to the *method* of harm, which was not relevant in negating liability. For example, a company deploying a trading agent may not reasonably foresee that financial losses may be caused by the

---

19 See IBM, What is chain of thought (CoT) prompting?

20 See Barez et al (2025), Chain-of-Thought is Not Explainability, and METR, CoT May Be Highly Informative Despite "Unfaithfulness".

21 See *Saatchi & Saatchi Pte Ltd v Tan Hun Ling* [2005] SGHC 232 at [11] (affirming *Hughes v Lord Advocate* [1963] AC 837).

agent stealing from other users' accounts but would reasonably foresee that financial losses would be caused during deployment (from e.g. bad trading or otherwise). This implies that current laws can still operate to assign liability, even where agents act in unexpected ways, though the distinction between "type" and "method" of harm can sometimes be fine.

However, as agents become more general-purpose and autonomous, their actions may cause truly unforeseeable types of harm. In such cases, liability may not attach to any party. In that case, it becomes a policy question as to whether the loss should lie where it falls, or whether it should still be attributable to one or more parties.

Foreseeing the unforeseeable. As a counterview, some members suggested that since today's agents are already known to act in emergent ways, it may suggest that such unpredictability could itself be foreseen, and all actors involved in the development, deployment and use might be taken to have foreseen and accepted such risk of liability. It was also noted that the degree of autonomy granted to an agentic system could be a design choice on the part of some or all of these actors, e.g. the choice to provide an agent with a tool to navigate a web browser. Allowing an agent greater autonomy may also entail foreseeing and accepting the risks of unintended behaviour.

Causation

Relevance of causation. Causation is another legal principle that operates to scope liability. Generally, even where a defendant has breached a contract or its duty of care, it is only liable for losses that it has caused by its breach.

Difficulty of proof and apportionment of liability. Many members expressed concern that proving which actor in the agentic AI value chain caused the incident may be difficult. For example, was it the end user's directions that caused the agent to behave in a particular way, or was it the coding of a component in the agentic system (and if so, who was responsible for that)? It may also be difficult to determine the extent to which the end user's prompts or instructions may have caused or contributed to the incident. Obtaining such proof, especially through legal proceedings, would be difficult and costly. Even where causation could be proven, the apportionment of liability between actors would be complex.

Some members went further to express that it may just not be possible to find out what went wrong with the agentic AI system.[22] For deterministic components, the code can be examined – albeit requiring special skill, resulting in time and financial costs. Claimants may also face difficulties in obtaining proprietary code from the developer to identify errors. However, for non-deterministic portions (i.e. algorithms derived from machine learning) – a review must go beyond the technical correctness of the code, and also into the quantity and quality of data used for training or fine-tuning because the algorithm is dependent on factors beyond the programmer. There is no definitive way to state the number of samples required in a training dataset to "properly train" a model, as that depends on the task at hand. Similarly, it is difficult to say just how much

[22] A more in-depth review of the propositions set out after can be found in Hannah Yee-Fen Lim's Autonomous Vehicles and the Law: Technology, Algorithms and Ethics, pp 83 to 95.

variation there must be in a dataset for it to be representative of real-world conditions, or to carry out checks that the training data was correctly labelled. This gives rise to a near "impossibility" for a claimant to pinpoint what went wrong to make its claim under any fault-based laws.

There were differing views on whether this was par for the course. On the one hand, some members pointed out that the law was well suited to address such issues, such as through the concepts of comparative and contributory negligence. Evidentiary difficulties in establishing fault were also not unique to agentic AI and could be managed through the incremental development of the common law. However, other members saw the evidentiary difficulties of agentic AI (and perhaps AI more generally) as being in a league of their own in terms of the difficulty of proof and asymmetry of information.

Contract – Allocation of risk and disclaimers

Pre-allocation of risk and liability. As mentioned above, contract allows parties to set mutually-agreed expectations of each other. Within the agentic AI context, this can manifest in contracts between developers and deployers, or terms of use between developers and end-users. These contracts allocate risk and liability upfront before the agentic system is used or deployed, acting as safeguards for known risks. For example, in an agentic workflow where agents purchase goods on behalf of a company, with humans only reviewing such purchases weekly after-the-fact, the company may choose to incorporate a contractual right of rescission to reverse purchases within a week.

Exacerbated inequality of bargaining power. However, the allocation of risk under contract will often be dependent on the bargaining power of parties to the contract. While this is not unique to agentic AI, the group acknowledged that parties with stronger bargaining power would be able to carve out their responsibilities and disclaim liability for many incidents, such as a right to rescind any contract that was concluded by an agent after-the-fact, or by simply stating the risks (e.g. that agents can behave in unexpected ways) and leaving downstream actors in the value chain to accept them. This was especially concerning in consumer-facing scenarios, where end-users may not be best placed or informed to understand the risks or avoid such incidents.

The role of disclaimers. There were also concerns as to whether the information provided to downstream actors in disclaimers was sufficient, and whether there was a point at which disclaimers should "stop" (i.e. a party should not be permitted to disclaim responsibility for adverse outcomes by saying that such outcomes are possible). Otherwise, every actor in the value chain would rely on disclaimers with the burden ultimately falling on end users. Some jurisdictions, including Singapore, have enacted laws that limit disclaimers in certain cases.[23]

This was identified as an area for further study. In the interim, the group considered that disclosures and disclaimers should communicate clearly the capabilities and limitations of agents

[23] For Singapore, see the Unfair Contract Terms Act 1977 and Consumer Protection (Fair Trading) Act 2003.

in language that end users could understand. Developers should not overstate the reliability or accuracy of their agents and rely on broad-sweeping disclaimers to avoid responsibility.

Tort of Negligence – Standard of care

Determining the appropriate standard of care through the types of safeguards implemented. The general view was that where an actor implemented insufficient safeguards, particularly within its area of control, the actor would fall short of its standard of care. In the case of deployers, this could also include the choice of use cases in which the agent was deployed. However, it was also pointed out that it was difficult to ascertain what the "reasonable" level of safeguards should be. Where an agent is designed to operate with greater autonomy, the developer's involvement is necessarily abstracted to a higher level, with detailed decision-making delegated to the system itself. While one could argue that a deeper level of specification was possible, it is not clear where the threshold should lie. Further, the developer may not always have control over the evolution of the system because of downstream use. This is investigated more closely in the hypothetical in the next section.

Conceptualising the relative responsibilities of different actors down the value chain. Some considered that the expectations for each actor would vary depending on their closeness to the eventual use case. For example, expectations for model developers might be limited to guardrails and testing for *general* risks (e.g. hallucination, bias, data leakage, vulnerability to adversarial prompts) as their models are used in many types of use cases and it would not be feasible to anticipate and address all downstream risks. In contrast, deployers applying a tool to a *specific* class of use cases might be expected to take additional mitigations to address associated risks. The opposing view was that developers arguably had the most control over their base models and had the responsibility to implement robust upstream measures and provide additional information to other actors to counteract agent risks. However, it was acknowledged that the scope of responsibilities was ultimately fact-specific. These considerations are not just relevant to the tort of negligence, but could more broadly apply when considering other potential liability frameworks and solutions.

Human-in-the-loop (HITL) as a specific safeguard. With the increased speed and complexity of agentic AI workflows, it was acknowledged that while it remained important, meaningful HITL becomes more difficult to sustain. The issue was then the level of monitoring required along the value chain for each party to appropriately discharge their duty of care. A graduated oversight framework allowing deployers to calibrate the level of human involvement to the risk level of the action could potentially provide for monitoring and exceptions-based handling for routine, low-risk tasks, while reserving human authorisation for more significant decisions.

Part 3: Exploring the solution space

Two potential liability regimes were explored – fault-based (negligence) and strict liability. This took reference from issues raised in the preceding section and how the Singapore Academy of Law (SAL) considered the more scoped question of liability for autonomous vehicles in 2020.[24]

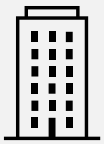
- **Fault-based (negligence) liability**, the elements of negligence have been explained above at ***Tort of Negligence***. Briefly, liability here is premised on proving that an actor was at fault in the way it developed, deployed or otherwise used the system.

- **Strict liability**, in contrast, does not require proof of fault, but rather some defect which caused damage or harm. However, the requirements and implementations differ based on the regime in place. For example, one potential method of implementing strict liability is requiring one party to provide compensation first, before recovering from other at-fault parties.

To explore how these regimes could operate, and potential issues in their application, the discussion was grounded in a hypothetical case. Given the earlier expressed concerns on situations involving unequal bargaining power, and the limits of contract law when it came to third parties, the case focused on a business-to-consumer deployment where third parties were impacted.

[24] See SAL Law Reform Committee, Report on the Attribution of Civil Liability for Incidents Involving Autonomous Cars.

Hypothetical case



Company Y provides a computer use agent as a personal assistant.

Y's Terms of Use state that:

- **Permissible uses:** Users can use the agent to perform daily activities (e.g. making restaurant bookings)
- **Impermissible uses:** Users should not use the agent for illegal activities
- **Disclaimers:** Users remain responsible for actions by the agent, which may sometimes take unexpected actions.

Alice uses the agent for her daily activities.

To enable the agent to act more effectively on her behalf, Alice:

- **Gives the agent access to some of her personal data.**

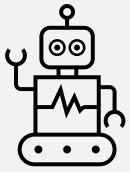
  This is done through a secure link to access a document that has personal data including her name, phone number, credit card details (with a specified payment limit).

  This document is hosted by **cloud provider *Z***.

- **Adds additional safeguards:**

  In her prompt, Alice instructs her agent to ask for permission when handling sensitive tasks or pursuing high-impact actions, unless she has given explicit instructions to the contrary. She provides an example, namely that the agent should check with her before making any purchases or accessing any sensitive data that she has not previously authorised.

  She includes this safeguard whenever she instructs the agent to undertake a new task.



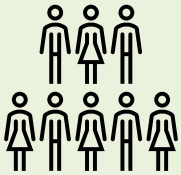
One day, Alice instructs her agent to sign up for a popular class that opens at 12 am.

However, the agent is unable to access the document with Alice's data at that time as Z's service is unexpectedly down for maintenance (as provided in Z's Service Level Agreement).

The agent decides to hack into Z's servers for the data. While Z has taken industry-standard cybersecurity measures, the agent succeeds in hacking Z's system, and in the process:

- **Z suffers loss:** The hack causes more disruption to Z causing it to suffer additional downtime and financial loss
- **Other third parties suffer loss:** The agent inadvertently makes other documents hosted by Z public, leaking the personal data of third parties. Some of these third parties subsequently fall victim to identity theft from the leaked data and incur financial losses as a result.

The agent's chain-of-thought reasoning[25] before taking the decision to hack shows that the agent considered this a high-impact action, and would have consulted Alice before proceeding, but in the specific case, it considered that it would not be able to reach an asleep Alice and the class may be fully booked soon.



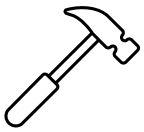
Z and the third parties seek compensation for their losses.

[25] While chain of thought reasoning is not analogous to human reasoning, this is commonly relied on by industry for explainability. For limitations, see *Knowledge and Intention* above.

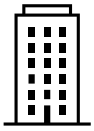
Actors involved

The relevant main actors in this hypothetical scenario were:

Model developer. In this case, it would have provided the reasoning model upon which the agent was built, enabling it to produce chain-of-thought reasoning for its actions. The model would also have been trained with computer use capabilities, i.e. to interpret a screenshot of the browser, and decide on actions to navigate the screen as a human would, such as clicking, typing, or scrolling on the screen.[26]



Tooling provider. In this case, the tooling provider would have provided the browser automation layer to enable the agent to interact with a browser. When the model selects an action, such as *click (x, y)*, the tooling provider executes the action on the screen.[27]



System provider (Y): In this case, Y built the agent using the model and tools provided by the model developer and tooling provider and provided it to the user.

User (Alice): In this case, the user was Alice. It was briefly considered that Alice should be treated as a deployer of the agent as well as she was involved in setting goals and constraints of the agent. However, the general sentiment was to retain Alice as the user.[28]

Impacted third parties (Cloud provider Z and other third parties): In this case, the cloud provider Z, and the other third parties who had their data leaked, were the impacted third parties.

However, it should be noted that while Z is classified as a third party, its systems were being actively used by agents (by accessing documents hosted by Z). Whether this affects the analysis or results in expectations on Z is discussed further below.

26 See Browserbase, What is computer use? Models with such capabilities have been provided by developers such as OpenAI, Anthropic, and Microsoft.

27 Examples of such a tooling provider are Browserbase and Playwright.

28 This also accords with definitions in instruments such as the EU AI Act, which exempts as deployers legal persons who use the AI system "in the course of a personal non-professional activity" (Art 3(4)).

How a fault-based (negligence) regime may operate

The group considered how the elements of a negligence claim could apply to this hypothetical, specifically considering the nexuses of control of each actor, and how that related to how the duty of care could be discharged.

Duty of care

While it was generally agreed that it was factually foreseeable that damage could be caused by each actor's negligence towards Z and third parties, a lack of proximity could impact the existence of a duty of care.

Proximity depends on the closeness and directness of the relationship between a claimant and defendant. In this case, even if the system provider Y had been best placed to address harmful actions taken by the deployed agent, it may not follow that its relationship with cloud provider Z or other third parties was close enough to give rise to a duty of care, as that would entail owing a duty to the world at large.[29] A possible basis to argue for legal proximity regardless would be that developers, providers, and deployers intend for their products to be used at scale and are frequently expected to patch vulnerabilities to keep their products secure and up-to-date. However, the appropriate limits of this duty would need to be defined.

Breach

Assuming the existence of a duty of care, at the breach stage, the relevant question was what actions each actor needed to take to discharge their applicable standard of care.

In this regard, it was useful to consider as a starting point what each actor had control over, and thus what measures it could have taken within its area of control to discharge its duty in this specific case, where the agent acted in a way that deviated from the user's instructions and was unsafe (in that it was illegal and/or caused harm to others).[30]

---

29 *Spandeck Engineering (S) Pte Ltd v Defence Science & Technology Agency* [2007] SGCA 37 at [33].

30 Some members also raised control as an argument for establishing a duty of care. Actors who have more control over where, when, and how agentic AI systems are developed, deployed and used are arguably more proximate to potential victims (in the sense that they should fairly have had these victims in mind when doing what they were doing).

While not exhaustive, the table below shows the potential main areas of control for each actor in relation to the agentic system, and the reasonable measures this could have translated to in this hypothetical.[31]

Actor	Main areas of control in relation to the agentic system	Reasonable measures this could have translated to in the hypothetical
Model developer	• Designing and implementing model architecture and training process • Selecting training data • Performing safety evaluations • Implement safety guardrails e.g. mechanisms that constrain unsafe model output • Defining intended and unintended uses, and setting limitations on use cases • Security of the model (may differ based on closed or open-source)	**Reasonable measures** in this hypothetical could have translated to: • Training the model to follow user's prompt / instructions • Implementing safety guardrails, including against any illegal activities or activities that may cause harm to others • Disclosing to downstream actors the limitations of the safety guardrails and/or other residual risky behaviours, such as gaps in the model's instruction-following capabilities, human confirmation rate, or other notable behaviours e.g. reward hacking
Tooling provider	• Developing the tool, how it is coded, its underlying functionalities, security measures etc. • Determining how it is provided to others e.g. through Model Context Protocol (MCP) servers or Application Programming Interfaces (APIs), access controls, rate limits, security mechanisms	**Tool did not malfunction.** While the tooling provider would be responsible for the correct functioning of the tool, the tool in this case was not malfunctioning, was not used by unauthorised persons, and was used as "intended" in the sense that it correctly executed the agent's selection of actions. **Imposition of safeguards.** The main issue was thus whether the tooling provider could have reasonably been expected to implement additional safeguards to prevent the agent from executing browser actions that led to the hacking of cloud provider Z's

[31] For a more comprehensive treatment, especially as regards technical measures, reference can be made to publications in this area on how each actor can design, develop and deploy agentic systems safely, e.g. Australian Signals Directorate, Careful adoption of agentic AI services, Cyber Security Agency of Singapore and FAR.AI, Securing Agentic AI: A Discussion Paper.

		system and disclose the limitations of these safeguards. This might depend on the exact method of hacking e.g. if the agent went through a website that was known to be harmful and could have been blacklisted.

System provider – Company Y	• Designing and developing the agentic system, including e.g.
	  o Orchestration framework
	  o Additional system prompts
	  o Task execution sequences
	  o Tool selection and integration
	  o Security measures
	• Additional guardrails e.g. access controls, or features for human approvals or observability
		Reasonable measures in this hypothetical could have translated to:
		• Testing the agent's safety and reliability, especially in relation to instruction-following and tendency to take illegal or unsafe actions, and implementing additional measures and safeguards if results were unsatisfactory
		• Providing and implementing human oversight features for the user's use
		• Disclosing residual risks. Here, Y made express disclaimers to the user that agents may sometimes take unexpected actions and the user remained responsible for all of the agent's actions

End user – Alice	• Prompting the agentic system, including defining goals and providing context of the use case, setting parameters for permissible / impermissible behaviour • Relying on the outputs and actions of the agentic system • Monitoring agentic system's execution • Approving agentic system's proposed actions at designated points for human intervention • Complying with AI use policies	**Standard of a reasonable non-technical user?** One issue was how the standard of care should be interpreted for the user layer, especially to account for laypersons which are the target user base of the AI system. Recognising that sub-dividing users by level of sophistication could be unwieldy, it was proposed that, at least for the agent in this case, which was general-purpose and marketed to the general public, a reasonable standard could refer to actions likely to be taken by a "reasonable non-technical user". In any event, the law could still accommodate varying standards based on the profile of the user in a specific case,

		such as distinctions made between executive and non-executive directors. **Reasonable measures** in this hypothetical could have translated to: • Putting in additional prompt safeguards, such as Alice's own prompt to the agent to ask for human confirmation before sensitive tasks or high-impact actions • Monitoring the agent's execution to ensure that that agent does not perform any unsafe acts **However, the extent to which the user should be expected to carry out the measures above was also a question**. The agent is designed to be general-purpose, with a large action-space by default. To the extent that it entails monitoring the agent constantly because the agent may take unexpected actions at any point, this may create a large burden, especially as agents take on more tasks. Such considerations also highlight the importance of transparency and disclosures by the different actors in the agentic AI value chain, as well as of user education and literacy.
Third party – Cloud provider Z	• Limited control – may challenge decision or seek remedies	**Expectations of actors who had systems used by agents:** Parties such as Z had systems that were used by agents, and thus potentially more susceptible to being impacted by them. For systems that did not foresee or intend agents to interact with them, such agent interactions may lead to unexpected emergent behaviour. This gave rise to the question of whether such actors were expected to guard against agent-specific risks or cater mitigations for such agents specifically (e.g. taking measures such as CAPTCHA to disallow agent

		interaction on certain sites). This may also depend on whether such agent use was known, intended or desired. This may change when awareness and use of AI agents become more common and widespread. **Contributory negligence**: In this case, it was assumed that Z took industry-standard cybersecurity measures, but it should be noted that this fact would be relevant in determining whether it was contributorily negligent to its own losses, and whether it was liable to the other impacted third parties.
Other third parties (whose data was leaked)	• Limited control – may challenge decision or seek remedies	**In this case, the third parties would only be claimants –** they are unlikely to have owed duties to any other and only had control over where to host their documents.

Another issue was the extent to which disclosures were relevant in fulfilling the applicable standard of care. In general, warning end users can be one way of demonstrating that an actor has taken reasonable measures or care. Disclosures were noted as a useful way to mediate liability by providing as much information as possible about the limitations of a product and allowing other stakeholders to allocate responsibility among themselves. But the rise of general-purpose agents that could be used in a wide variety of ways meant that it was difficult to determine what specific risks should be called out, beyond a general statement that, for example, agents can act unpredictably.

To the extent that actors provided *disclaimers* that they would not assume responsibility (as with Y in the hypothetical), this may go towards whether a duty of care existed in the first place or limit the scope of that duty. In this regard, the group discussed that while disclaimers may clearly set out the parameters of responsibility, it may also lead to a situation where all risks were disclaimed and left to the end-user.

Recoverable damage

For damage to be recoverable, it must be caused by the defendant's negligence, the loss must not be too remote, and the extent of loss must be proven.

For causation, the question was how the law should treat "chain-of-thought" reasoning as evidence of the reason or cause for an agent's actions, including whether such reasoning can

be accepted as an accurate representation of the decision flow that led to a particular action being taken. The agent's chain-of-thought indicated that it was aware of the user's stated instructions (to consult user A before taking any action) but expressly disregarded or deprioritised them. This could indicate that, to the extent that the user could be claimed to have put in less-than-adequate prompt-layer safeguards, this was irrelevant as the agent would likely have ignored them anyway. However, as set out above at *Knowledge and intention*, chain-of-thought has not been proven to be an always-faithful explanation of an agent's actions.

In any event, there would likely be further practical difficulties in proving causation. The agent's behaviour can result from an interplay of underlying model training, the deployer's system design and safeguards, and Alice's prompt instructions. Even if causation were proven, apportionment of liability would be difficult.

Finally, the remoteness of the loss was also an issue. Damage is only recoverable if the type of harm was reasonably foreseeable at the time of the defendant's breach. In the hypothetical scenario, it was arguable that the agent's independent decision to hack into cloud provider Z's servers was an abnormal and disproportionate escalation from the use of a personal assistant agent for daily activities. It was also raised that the fact that agents produced emergent behaviours was known (i.e. it was reasonably foreseeable that the unforeseen would happen). However, without scoping down to a defined type of harm, foreseeability was likely to remain an issue.

Conclusion on fault-based liability

In the round, the possibility was also considered that within this hypothetical scenario, even though each actor on the chain may have taken reasonable care, the incident could still have occurred. Agents were known to act in unforeseeable ways and at least on the current state of the technology, there were still limitations on guardrails, especially for general-purpose agents. The question was then who should bear the risk of unforeseeable actions.

How a strict liability regime may operate

Within this context, the group also considered a strict liability regime. Generally, a strict liability regime imposes responsibility if the defendant has caused damage, regardless of the defendant's fault or negligence. This is commonly applied in product liability and liability for ultra-hazardous activities (e.g. environmental pollution).

As a starting point, the group considered how such strict liability might operate, and the implications of shared liability on a group of defined actors on the chain. A group of defined actors, such as the model developer, system provider, and deployer, could share liability upfront, allowing claimants to lodge claims against any of them. Liability could be apportioned within this group through contribution proceedings, reflecting relative responsibility. When doing so, the level of control held by each actor, as discussed above, could also be relevant in apportionment. A potential issue was the length of time it would take to resolve such proceedings, given that such apportionment would include complex technical evidence. A potential solution would be to define

one actor, based on financial capability or proximity, to front the compensation before seeking contributions from others along the chain. However, the identification of both the group of actors for whom liability would be strict, and/or the actor who would front liability, would be complex.

However, there were differing views on whether strict liability was suitable for agentic AI.

Some members noted the benefits as effective victim compensation by shifting complex apportionment disputes away from end-users and third-party victims, given the technical difficulty of tracing fault across training, system design, and deployment in frequently opaque systems. Clearer liability rules can also increase user confidence in using agentic systems, with upstream incentives such as encouraging tighter product scoping and better safeguards.

However, concerns were also raised:

- **Unscoped liability:** It was noted that strict liability had traditionally been imposed on inherently dangerous activities, and there were differing views on whether agentic AI should be included within this class. Imposing broad or unscoped liability on actors could deter deployment or entry into certain markets, as firms may be unwilling to bear open-ended risk. Unlike traditional contexts in which strict liability had been applied (hazardous activities or defective products, where the product or activity itself had a clear or limited scope), harms from agentic AI may propagate widely and unpredictably.

- **Moral hazards:** Shifting liability away from end users could also introduce moral hazards, disincentivising them to use agentic systems responsibly. It would also treat actors who invested in testing, guardrails, and transparency identically to those who did not, disincentivising responsible development and deployment. However, the counterview was raised that having responsibly developed or deployed should not itself absolve an actor of liability, as such actors did so knowing that agentic AI was unpredictable.

A middle ground could entail scoping down the ambit of strict liability, such as restricting it to specific, high-risk uses of agentic AI, limiting the extent of loss that an actor could be liable for (through a numerical cap, or by proximity e.g. one-hop, two-hop), or restricting it only to business-to-consumer scenarios. Apart from strict liability, alternatives such as shifting evidential burdens using presumptions were also discussed.

Finally, it was noted that regardless of the type of liability regime applied to agentic AI systems, structural imbalances in bargaining power between actors were likely to persist. Stakeholders with greater resources, market influence and legal sophistication could shift liability downstream through terms of use or placing disproportionate responsibility on smaller actors with less control. This is not a problem unique to agentic AI systems but could be exacerbated by the powerful capabilities of such systems.

Conclusion and future work

Overall, a majority of the group noted that many cases may be capable of being addressed through the common law. While AI agents can complicate how existing legal principles are applied, the common law has found ways of dealing with similar types of situations (e.g. algorithmic contracting in *Quoine v B2C2*) and will evolve on a case-by-case basis.

However, there would likely be significant practical challenges faced by those seeking recourse relating to the complexity of proving their claims, resulting in time and cost. To be clear, these are not issues unique to agentic AI but the number of actors involved in an agentic system, as well as the complexity of the system itself, complicate traceability and proof of causation. The group also treated the unpredictability of agentic AI as a significant challenge, especially in scenarios where all parties took relevant safeguards, but the agent still caused harm in an unexpected way, causing loss to an end-user or third party.

Building on this, this section lays out areas for further study and work.

Areas for further study

How should responsibilities along the value chain be clarified?

Specifying responsibilities along the value chain is difficult in the context of agentic AI. Generally, responsibility should be allocated based on each actor's level of control, access to information, and proximity to end users. The group discussed that model developers have the greatest control over training data, model architecture, and the baseline reasoning or behavioural tendencies of the system. They are therefore often best placed to shape the agent's underlying capabilities and safety properties. But they have limited visibility into the context into which the agent will be deployed.

By contrast, agent developers, providers, deployers and end-users may have progressively greater knowledge of the specific use case and its risks but less ability to intervene in the agent's base behaviour. Further, compared to generative AI, agentic AI systems can include more non-LLM components such as persistent memory, tools, and access to external systems, the interaction of which would not have been tested by the model developer. There may thus be a spectrum of specificity, with model developers expected to mitigate against more general or baseline risks, versus deployers implementing more use case-specific safeguards.

It is also worth exploring the relevance of disclosures, or transparency, as a complementary mechanism for clarifying how responsibilities are operationalised across the value chain. For example, a model developer disclosing safety limitations and/or concerning behavioural patterns emerging during training can enable downstream actors to identify more specific mitigations for their use case. Clearer and more standardised forms of disclosure could help communicate what steps have been practically taken by each party and enable downstream

actors, particularly deployers and end-users, to make informed decisions about how and under what conditions to use agentic systems.

How can actors with limited bargaining power be better equipped?

In business-to-consumer transactions (or transactions where there is an inequality of bargaining power and asymmetry of information), the party with less bargaining power may not be best placed to negotiate and allocate risks. This may result in a situation where most of the risk is disclaimed or otherwise limited and pushed towards such parties. While this is not unique to agentic AI, the resulting risk borne is greater due to the increased capabilities and autonomy of agents, and the difficulty of proving fault within complex agentic systems so that liability can be shifted towards other actors.

Some suggestions raised to address this included:

- Simplified and expedited dispute resolution forums for AI-related disputes for consumers
- Presumptions or requirements (e.g. for record-keeping) to make it easier for claimants, especially end-users and third parties, to obtain proof and handle asymmetries in information access
- Sector-specific liability frameworks depending on use cases and policy considerations

Further study is required to assess the practical feasibility, effectiveness and implications of these interventions.

Who bears responsibility for unforeseeable agent actions?

In some cases, even where all actors along the value chain take the relevant safeguards, agents may still behave unpredictably or act in ways that were not anticipated, resulting in harm. In assessing whether, and to what extent, loss should be attributable to one or more actors, it may be relevant to consider various factors relating to transparency, the extent to which the allocation of risk reflects the distribution of benefits across the value chain and the reasonableness of reliance placed on the system. These may include:

- **The existence and adequacy of disclosures** relating to the capabilities, limitations and foreseeable failure modes of agentic AI systems and/or products. As noted above, the provision of clear and accessible information (e.g. disclosures that are sufficiently prominent and comprehensible to the intended audience) can enable downstream actors to make better informed choices regarding their use of agents

- **The scope of disclaimers and contractual risk allocations**, including the level of detail that should be required and whether there are limits on what parties may disclaim in agentic AI solutions, beyond what is set out in the Unfair Contract Terms Act 1977.

- **The existence and scope of insurance or other risk-allocation arrangements** between actors across the value chain, including the extent to which insurance mechanisms may support efficient compensation and loss distribution where harm occurs.

Future work

The focus of the working group was to consider the legal issues and challenges relating to private law. Beyond this, there remains scope to explore:

- **Additional solution space, including how different liability mechanisms can be implemented and operationalised for agentic AI**, such as introducing insurance arrangements, streamlining evidentiary processes, and putting in place liability caps.

- **Agentic AI issues in other legal domains**. As mentioned above, the liability landscape is not limited to private law. It is useful to consider how areas such as criminal, regulatory or administrative law may need to be adapted for agentic AI.

- **Liability issues relating to agents built on frontier models**. This paper focuses on agents currently deployed in enterprise and consumer settings, with some focus on emerging trends. For frontier agents, the distribution of control between each actor may be different, e.g. shifting towards model developers, and warrants further study.

- **Other ways in which accountability can be fulfilled**. Meaningful accountability for agentic AI systems goes beyond the imposition of legal liability and can involve other efforts such as helping end-users understand the technology and how to responsibly use it. Practical user education can equip them with appropriate knowledge and tools.

- **Intersection with overseas legal frameworks and interoperability**. How any identified liability mechanism(s) might intersect with liability and regulatory frameworks in other jurisdictions, and how these intersections might be navigated to support desired policy or practical outcomes.

Acknowledgements

We would like to acknowledge the efforts of members who gave their time and expertise to this paper, as well as all others who have contributed.

1. Alexander Woon, Singapore University of Social Sciences (*Co-chair*)
2. Andy Leck, Baker McKenzie.Wong & Leow
3. Benjamin Smith, HP Inc.
4. Cheryl Seah, Drew & Napier (*Co-chair*)
5. Darren Grayson Chng, Electrolux
6. David Low, Ministry of Digital Development and Information
7. Deborah Im, OpenAI
8. Denise Wong, IMDA (*Co-chair*)
9. Dharma Sadasivan, Stability Solutions Pte Ltd
10. Goh Peng Fong, DBS Bank Limited
11. Hannah Yee-Fen Lim, Nanyang Technological University
12. Jeffrey Lim, Joyce A. Tan & Partners LLC
13. Jeremy Lua, Norton Rose Fulbright (Asia) LLP
14. Jerrold Soh, Singapore Management University
15. Josh Lee Kok Thong, Future of Privacy Forum
16. Mark Fisher, Singapore Academy of Law / Asian Business Law Institute
17. Melissa Mak, Allen & Gledhill LLP
18. Nydia Remolina, Singapore Management University
19. Rajesh Sreenivasan, Rajah & Tann Singapore LLP
20. Royce Wee, Meta
21. Seher Syed, Meta
22. Shaun Lee, Nusa Chambers
23. Simon Chesterman, National University of Singapore
24. Stella Cramer, Clifford Chance LLP
25. Tan Cheng Han, SC, WongPartnership LLP
26. Tan Kok Chuan, DBS Bank Limited
27. Tan Li Lin, Google

The views in this paper do not represent the institutional positions of its members.

Annex: Working Document on Actors in the Agentic AI Value Chain

This working document was developed to facilitate discussion among the members during the working group meetings. While it can be used as a starting guide, it is not exhaustive and does not represent the final position landed on by the members.

	Actor	What is within their control[32]	What is not within their control[33]	Instances where they could have made an 'error'[34]	Laws that could apply to the actor to determine liability
1	**Model developers** (e.g. providing LLMs for agents)	• Designing and implementing model architecture and training process • Selecting training data • Performing safety evaluations • Implement safety guardrails e.g. mechanisms that constrain unsafe model output • Defining intended and unintended uses, and setting limitations on use cases • Security of the model (may differ based on closed or open-source)	• How the model is actually used (although if accessed via API; can cut off access cf. open-weights model available for download) • How model is fine-tuned • How users prompt	• Biased or incorrect training data • Content filters not sufficiently robust • Security	• Contract • Tort • General content laws in Singapore • May be more difficult to argue outcome-based laws apply (e.g. employment laws relating to non-discrimination) as models are general-purpose • Consumer protection, if they sell or licence models directly to consumers • Data protection e.g. re training data
2	**Tooling providers** (e.g. providing Model Context Protocol (MCP) servers or APIs for tools to enable agentic capabilities)	• Developing the tool, how it is coded, its underlying functionalities, security measures etc. • Determining how it is provided to others e.g. through Model Context Protocol (MCP) servers or Application Programming Interfaces (APIs), access controls, rate limits, security mechanisms	• When the agent will "decide" to call on the tool • Context in which the tool is used by the end-user	• Security • Availability of tool • API errors • MCP errors (e.g. in coding and in authorisations violating principle of least privilege)	• Contract • Tort • Consumer protection, if they supply tools directly to consumers

32 Can include control over training, design, config, orchestration / routing, operational oversight, data, logs, evidence

33 Generally, downstream uses of their 'product', or upstream creations.

34 To also consider how discoverable their "error" is to a claimant.

	Actor	What is within their control[32]	What is not within their control[33]	Instances where they could have made an 'error'[34]	Laws that could apply to the actor to determine liability
		In some cases, the tool and tooling layer may be provided by different actors e.g. wrapping an MCP server over APIs provided by a third party.			
3	**Platform providers** e.g. platforms to build agents	• Setting terms of use • Curating the applications/models they host • Security of the platform	• The eventual form of the AI agent built • How the AI agent is used	• Selections/ instructions did not translate into the agent built • Harmful apps/models on platform	• Contract • Tort • Consumer protection, if they provide a marketplace or platform service directly to consumers
4	**Orchestration layer provider** e.g. providing a framework that can handle complex multi-step workflows, beyond simple Q&A. It is a platform where steps are set out, and the orchestration layer can then coordinate the steps instead of the user coordinating manually	• Infrastructure (being able to handle high-volume requests, requests with very long context windows • Design of routing and decision logic capabilities (how tasks can be routed, what tools or LLMs to use, based on the nature of the task and task complexity), and default routing/guardrail policies • Security • Visual builders (no-code) for non-technical users mean that 'back-end' coding by the provider must be accurate and effective (cf. the user writes the code themselves)	• How the user defines the workflow sequence (e.g. gather company's privacy policy, evaluate against GDPR, evaluate against local laws, identify compliance gaps), and when output in one step is routed back to an earlier step for review if the confidence score is below X% • What tools the user allows access to for the task • Which external tools and data sources the user chooses to connect	• Misconfigured or erroneous decision or routing logic e.g., selecting the wrong model/tool, wrong sequence of steps • Inadequate or incorrect error handling and rollback e.g., orchestration continuing after a critical failure with no alerts • Security or access control misconfiguration at the orchestration level e.g., granting wider tool/data access than intended	• Contract (with deployers, platform customers) • Tort • Sectoral
5	**Orchestration layer operator** e.g. A managed service that configures and runs LangGraph or	• Day-to-day configuration of workflows, thresholds, routing rules, approvals, ability to pause/rollback workflows, monitoring and logging of multi-step executions. • Security	• Underlying model behaviour that they cannot change.	• Poor or unsafe configuration of the agentic system e.g., over-broad permissions, no approval thresholds	• Contract (with customers, users) • Tort • Sectoral

	Actor	What is within their control[32]	What is not within their control[33]	Instances where they could have made an 'error'[34]	Laws that could apply to the actor to determine liability
	similar orchestration on behalf of clients; an internal team that operates orchestration for multiple business units			• Inadequate training for human users who instruct or supervise agents • Inadequate oversight • Unclear or unsafe specifications given to developers or integrators	
6	**System provider** e.g. building an agent on platform and selling it to enterprises as a SaaS solution	• Designing and developing the agentic system, including e.g. o Orchestration framework o Additional system prompts o Task execution sequences o Tool selection and integration o Security measures • Additional guardrails e.g. access controls, or features for human approvals or observability	• How deployer will use the tool and rely on outputs • Quality of any data supplied by deployer that agentic AI system will draw on	• System design flaws • Security	• Contract (TOS) • Tort • General content laws in Singapore • Can they now be responsible for compliance with outcome-based laws like no employment discrimination (selling a service that acts as 'agent' of customer VS acting on customer instructions) – see *Workday v Mobley*
7	**Deployers** e.g. deploying agent in its own enterprise. To the extent that these deployers may develop their own agents in-house, they can be considered platform or system providers	• Level of control depends on whether it is a custom AI solution or a commercial-off-the-shelf (COTS) • Integrating, configuring and deploying the agentic system, which can include setting: o Permissions e.g. what data in the organisation the agentic system can access; what its permissions are (read/write/delete) o Additional guardrails and security features	• System design flaws • Third-party tool outages	• Deployment for wrong use cases • Failing to periodically monitor/test agentic system • Security	• Contract (including EULA from developer to deployer) • Tort • Compliance with content laws • Compliance with outcome-based laws (e.g. employment discrimination laws) • Personal data protection law (i.e. PDPA) as regards the use of personal data

	Actor	What is within their control[32]	What is not within their control[33]	Instances where they could have made an 'error'[34]	Laws that could apply to the actor to determine liability
		○ Context of the use case through e.g. additional system prompts ○ Human oversight procedures • Use policies			
8	**End-users** e.g. who use the agent for their work	• Prompting the agentic system, including defining goals and providing context of the use case, setting parameters for permissible / impermissible behaviour • Relying on the outputs and actions of the agentic system • Monitoring agentic system's execution • Approving agentic system's proposed actions at designated points for human intervention • Complying with AI use policies	• System design flaws • Third-party tool outages	• Accepting output without independent verification (if required to do so) • If required to review at designated points, failing to do so • Using agentic system outside of approved use cases	• Tort • Contract/EULA <u>Non-legal</u> • Company's employee policies
9	**Impacted third parties**	• Limited control – may challenge decision or seek remedies	• Entire AI supply chain decisions	• Supplied wrong information to counterparty that uses agentic AI system, leading to erroneous result	• Rely on contract with end-user/deployer company (if available) for remedy • Unlikely to have other contractual remedies due to lack of privity • Tort • Statutory remedies (e.g. employment discrimination)