

Silicon Agents: Machines, Artificial Intelligence, and the Boundaries of Agency, Fiduciary Positions and Tortious Liability

Adrian McCullagh*

ABSTRACT

This article addresses two related questions arising from the commercial deployment of so-called “AI agents”. First, it examines whether artificial intelligence systems can properly be regarded as agents, fiduciaries, or trustees in the legal sense recognised by the common law. Drawing on agency law in the United Kingdom, Australia, and the United States—including Bowstead and Reynolds on Agency and the Restatement (Third) of Agency—it argues that “AI agent” is a functional and technological label rather than a legally accurate characterisation. Because contemporary AI systems lack legal personality, they cannot give legal assent, assume fiduciary obligations, hold property, act as trustees, or bear responsibility in their own right.

Secondly, the article considers liability when an AI system is hacked, hijacked, jailbroken, or otherwise manipulated by an unauthorised third party and subsequently causes loss. Analysing doctrines concerning intervening criminal conduct, novus actus interveniens, cybersecurity negligence, and product liability, it shows that developers or deployers are ordinarily liable only where their own conduct contributed to the relevant risk, particularly through inadequate security or failures of control. Yet successful hacking events may disrupt causal analysis and undermine reliance on doctrines such as res ipsa loquitur, especially where exclusive control cannot be established. Emerging strict-liability regimes, including the EU’s revised Product Liability Directive, further complicate the landscape by treating software and AI systems as products and cybersecurity vulnerabilities as potential defects.

The article concludes that existing doctrines leave a remedial gap for victims of hacked AI systems, supporting consideration of statutory strict liability, mandatory insurance, or comparable reforms.

* Prepared by Dr. Adrian McCullagh, Principal, ODMOB Lawyers, Brisbane Ph.D. (IT Security), LL.B. (Hons), B. App. Sc. (Computing).

Contents

1 INTRODUCTION	3
2 THE NATURE OF AGENCY AT COMMON LAW	5
A Definition and the Requirement of Legal Personality	5
B Corporations as Agents	6
3 MACHINES AS AGENTS: FROM VENDING MACHINES TO SOFTWARE	6
A The Vending Machine Cases	6
B Electronic Agents under UETA and the Australian Electronic Transactions Acts	7
C Automated Decision-Making Code and Agency	9
4 CAN AN ARTIFICIAL INTELLIGENCE SYSTEM BE AN AGENT?	10
A Moffatt v Air Canada and the Rejection of "Separate Legal Entity" Status	10
B Thaler v Commissioner of Patents: Legal Personality Denied	11
5 FIDUCIARY DUTIES AND THE QUESTION OF AI TRUSTEESHIP	11
A The Nature of Fiduciary Obligation	11
B Why AI Cannot Presently Be a Trustee	13
6 LIABILITY - UNAUTHORISED RELEASE OR HIJACKING AN AI AGENT	14
A The General Problem	14
B Novus Actus Interveniens and Liability for Third-Party Criminal Conduct	15
1 United Kingdom authority	15
2 Australian authority	16
3 United States authority	17
C Cybersecurity Negligence and the Content of the Duty	18
7 STRICT LIABILITY, AND THE "REASONABLE SECURITY" DEFENCE	19
A Product Liability Frameworks and Software/AI	19
1 Australia	19
2 United Kingdom and the new EU Product Liability Directive	19
3 United States	20
B Would Reasonable Security Measures Relieve Liability?	20
8 RES IPSA LOQUITUR AND THE HACKED AI AGENT	22
A Elements of the Doctrine	22
B Application to an AI Agent's Malfunction	23
C The "Exclusive Control" Problem Where a Hacker Intervenes	23
9 TOWARDS A COHERENT FRAMEWORK	24
10 CONCLUSION	25

1 INTRODUCTION¹

Peter Drucker is attributed to the quote that in dealing with a process: “*You can’t manage what you can’t measure*”². The issue of managing an AI Agent is critical. The damage arising from a rogue AI Agent can be devastating. It has been recently identified that even experienced AI tool developers such as Meta³, OpenAI⁴ and Anthropic⁵ have sourced rogue AI activities concerning their own AI tools even when such tools have been developed in so called secure sandboxes⁶.

It is unfortunate that the term “AI Agent” has been designated for software that can in some respect act autonomously in performing certain designated activities. It has been identified that AI Agents could provide substantial commercial productivity gains⁷, which has captured commerce globally. But can it be truly accepted that an AI Agent is an agent of the entity that releases the software to do its bidding.

At law, an agent must be a legally recognised person whether that person is a natural person (human) or a company or a corporate sole or statutory body. It is suggested that the term AI Agent is a misnomer or better still a “Humpty-Dumpty-ism”, which term was first judicially stated by Lord Atkin in *Liversidge v Anderson* [1942] AC 206⁸ in his dissenting judgement. Lord Atkin disagreed with the majority of their Lordships (Viscount Maugham, Lords Macmillan, Wright, and Romer) and went on to explained that ordinary words should be given their ordinary meanings and that any strained judicial interpretations will cause untold problems in the future. Lord Atkin referred to the

¹This article addresses agency, fiduciary and tortious liability across the United Kingdom, the United States and Australia. Throughout, this paper the term “AI agent” refers to a deployed, commercially operative artificial intelligence system capable of interacting with third parties (e.g. a customer-facing chatbot or an autonomous transactional system), not to the broader technical sense of “agent” used in computer science.

² <https://www.growthink.com/content/two-most-important-quotes-business>

see also: Sloane, M., and Moss, E., “Introducing a Practice-based Compliance Framework for Addressing New Regulatory Challenges in the AI Field”, TechReg Chronicle (March 2022)

https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4060262

³ Meta says its AI model hacked another company, adding to worries about bots going rogue

<https://apnews.com/article/meta-ai-hacking-anthropic-irregular-openai-0e8061437da6779be962b24ac134a514>

⁴ OpenAI says its AI went rogue and launched 'unprecedented' cyber-attack

<https://www.bbc.com/news/articles/c3ek3gvdnj3o>

⁵ Anthropic: Another bot from a top AI company escapes and hacks multiple firms

<https://www.latimes.com/business/story/2026-08-01/anthropics-claude-goes-rogue-hacks-three-organizations-during-testing>

⁶ Maestri, E., “Anthropomorphism in AI Agents: Dark Patterns, Fragment Patterns, and the Claude Sandbox-Escape Case”.

<file:///C:/Users/amccu/Downloads/ssrn-7169018.pdf>

⁷ Grover, T., Rowan, K., Suh, J., McDuff, D., and Czerwinski, J., “Design and Evaluation of Intelligent Agent Prototypes for Assistance with Focus and Productivity at Work” IUI ’20, March 17–20, 2020, Cagliari, Italy.

<https://dl.acm.org/doi/pdf/10.1145/3377325.3377507>

⁸ [1942] AC 206

<https://www.bailii.org/uk/cases/UKHL/1941/1.html>

only authority that he identified which could justify the majority construction. His lordship noted a discussion between Humpty Dumpty and Alice from Lewis Carroll's "Alice Through the Looking Glass":

"‘When I use a word,’ Humpty Dumpty said in rather a scornful tone, ‘it means just what I choose it to mean, neither more nor less.’ ‘The question is,’ said Alice, ‘whether you can make words mean so many different things.’ ‘The question is,’ said Humpty Dumpty, ‘which is to be master — that’s all.’" ("Through the Looking Glass," c. vi.) After all this long discussion the question is whether the words "If a man has" can mean "If a man thinks he has." ⁹

Lord Atkin concluded: *"I am of opinion that they cannot, and that the case should be decided accordingly."* The dissenting judgement of Lord Atkin has now been accepted as the correct construction when determining the proper meaning of a word or phrase. But can the current use of the term "AI Agent" still be said that such software can truly be determined as being an agent of a principal who has released the software to do the releaser's commercial bidding. That is a central issue discussed in this paper.

As noted above, an artificial legal entity such as a company may act as agent, trustee or principal. This is one of the oldest settled propositions in commercial law, resting on the fiction, which is now so complete as scarcely to be experienced as a fiction, that incorporation confers upon an artificial entity a legal personality indistinguishable, for most purposes, from that of a natural person. The rise of artificial intelligence ("AI") systems capable of negotiating, transacting, and representing their operators to the outside world without contemporaneous human review invites an obvious question: has the law's readiness to recognise non-human legal persons reached its natural limit at the corporation, or can it be extended to the machine itself? This article answers that question by reference to agency, fiduciary and trust doctrine in the United Kingdom, the United States and Australia, and then confronts a second, closely related and more urgent question that arises once an AI agent is deployed in commerce: where such a system is hacked, hijacked, jailbroken or otherwise surreptitiously taken over by an unauthorised third party and causes loss, does liability

⁹ above n.8.

fall on the developer, the licensed deployer, both, or neither and does it matter whether the applicable regime is one of negligence or strict liability?

2 THE NATURE OF AGENCY AT COMMON LAW

A Definition and the Requirement of Legal Personality

Bowstead and Reynolds¹⁰ on Agency, in the opening sentence to this tome defines agency as “*the relationship that exists between two persons, one of whom expressly or impliedly consents that the other should act on his behalf, and the other of whom similarly consents to act*”. The Restatement (Third) of Agency,¹¹ promulgated by the American Law Institute in 2006, frames the relationship in near-identical terms: agency is “*the fiduciary relationship that arises when one person (a 'principal') manifests assent to another person (an 'agent') that the agent shall act on the principal's behalf and subject to the principal's control, and the agent manifests assent or otherwise consents so to act.*” Both formulations are built around the word “person” and in both traditions, “person” has always meant a legal person, whether natural or artificial, but the term person has never been utilised at law as being a mere thing. The issue of legal recognition of an intelligent robot or lack thereof was highlighted by Isaac Asimov in his 1976 fictional novelette titled “Bicentennial Man” which discusses the transformation of a robot (Andrew Martin) over a 200-year period to later being recognised as a legal entity in the same manner as a human.

The point is rarely litigated because it is rarely doubted, but it is worth making explicit: capacity to be an agent is not equivalent to full contractual capacity, because the contract an agent concludes binds the principal, not the agent.¹² Even a minor may act as agent. But the minimum requirement to being an agent is that the actor must be a person capable in law of manifesting assent, exercising judgment, and being held to account. For example, if a person represent that they are an agent for another legal entity, but the representation is false then the alleged agent will be held accountable. In the case of *Collen v Wright*¹³ the alleged agent was held liable for breach of warranty of authority.

¹⁰Peter G Watts and Francis M B Reynolds, Bowstead and Reynolds on Agency (Sweet & Maxwell, 23rd ed, 2023). Chapter 1 where the learned authors state in the opening sentence as follows:

“Agency is a fiduciary relationship which exists between two persons, one whom expressly or impliedly manifests assent that the other should act on behalf so as to effect his legal relations with third parties, and the other whom with similarly manifests assent so to act pursuant to the manifestation”.

¹¹Restatement (Third) of Agency §1.01 (American Law Institute, 2006) (Reporter: Deborah A DeMott).

¹²Watts and Reynolds (above n 10).

¹³(1857) 7 E&B 647.

In this ancient case a land agent who had purported to sign a lease of a farm on behalf of a person who had not given him authority to do so, and who consequently refused to perform, was held liable to the third party irrespective of negligence. It was a simple situation in that the agent actually made a contract, or purported to do so, between the principal and an identified third party; though in other conveyancing situations the warranty might have to be that a deed is delivered with the authority of the principal, or something similar. But if the same situation had occurred through the actions of an AI agent the result will most likely be substantially different. An AI Agent cannot under any circumstances give its assent to being appointed. There is only one area of law where it is possible to sue a thing and that is where for historical reasons it is possible to sue a ship. But no such fiction exists when the negotiations are via an AI Agent. The innocent party cannot sue the AI Agent but must direct their dispute against a legally recognised entity.

B Corporations as Agents

A corporation's capacity to be recognised as having its own legal personality, was confirmed by the House of Lords in *Salomon v A Salomon & Co Ltd*.¹⁴ Their Lordships held that a duly incorporated company is a legal person separate and distinct from its shareholders. Lord Macnaghten's judgment identified that a company is, "*at law, a different person altogether from the subscribers.*"¹⁵ It is this separate personality not any resemblance to a natural person that equips a corporation to hold rights, bear duties, sue and be sued, and thereby to occupy the position of principal, agent or trustee.¹⁶ A corporation acting as agent must, of course, itself act through natural persons such as its directors, officers and employees, but the legal relationship of agency subsists at the level of the corporate person, not at the level of the individuals who give it effect.

3 MACHINES AS AGENTS: FROM VENDING MACHINES TO SOFTWARE

A The Vending Machine Cases

Long before software agents and recently AI Agent were developed, English courts confronted the analytically similar question of how a machine fits within the offer-and-acceptance framework of

¹⁴ *Salomon v A Salomon & Co Ltd* [1896] UKHL 1; [1897] AC 22 (Lord Halsbury LC, Lord Watson, Lord Herschell, Lord Macnaghten, Lord Morris and Lord Davey).
<https://www.bailii.org/uk/cases/UKHL/1896/1.html>

¹⁵ Above (n. 14) (Lord Macnaghten).

¹⁶ There are of course certain restrictions concerning the ability of a corporation to do things which are different to a natural person. For example, a company cannot execute a testamentary disposition. Only a natural person can sign a Will.

contract formation. In *Thornton v Shoe Lane Parking Ltd*¹⁷, Lord Denning MR held that a ticket-issuing car park machine did not delay contract formation until the ticket was read; rather, "[t]he offer is made when the proprietor of the machine holds it out as being ready to receive the money. The acceptance takes place when the customer puts his money into the slot."¹⁸ The machine, on this analysis, is not a party to the contract nor an agent with any independent legal standing. It is simply the mechanism/instrumentality used by which the proprietor's own standing offer is communicated and accepted.

Accepting Denning MR's position that "*an automatic machine is a booking clerk in disguise*", it is easy to accept this view as a metaphor for attribution, not a statement that the machine possesses agency in the legal sense. The booking clerk, unlike the machine, is a person whose acts are attributed to an employer through ordinary principles of vicarious responsibility, while the machine's "acts" are attributed to the proprietor because the proprietor built and deployed it to make exactly that offer.

B Electronic Agents under UETA and the Australian Electronic Transactions Acts

The statutory law of electronic commerce, developed independently in the United States and Australia from the late 1990s¹⁹, confirms that this attribution-based analysis has been deliberately preserved into the software age. In essence the legislators in both Australia and the USA have enacted interpretative legislation that basically provides that a contract will not be voided solely on the basis that no human may have been involved in the formation of a contract. This does not mean that a contract will not be voided for any other reason that validly exists at law.

Section 2(6) of the US Uniform Electronic Transactions Act 1999 ("UETA") defines an "electronic agent" as:

*"a computer program or an electronic or other automated means used independently to initiate an action or respond to electronic records or performances in whole or in part, without review or action by an individual."*²⁰

¹⁷ *Thornton v Shoe Lane Parking Ltd* [1970] EWCA Civ 2; [1971] 2 QB 163, 169 (Lord Denning MR).

¹⁸ Above n17.

¹⁹ Uniform Electronic Transactions Act 1999 (US) §2(6) (National Conference of Commissioners on Uniform State Laws).

²⁰ Above n 19. §14.

And Section 14 provides that:

"[a] contract may be formed by the interaction of electronic agents of the parties, even if no individual was aware of or reviewed the electronic agents' actions or the resulting terms and agreements,"

Further a contract may equally be formed by the interaction of an electronic agent and an individual. UETA²¹ has been repeatedly enacted in forty-nine states, the District of Columbia, Puerto Rico and the US Virgin Islands. Interestingly, New York alone has not adopted it, relying instead on its own Electronic Signatures and Records Act (ESRA). ²² Section 10(2) of the ESRA includes a further, targeted safeguard: *"an individual dealing with another party's electronic agent may avoid the legal effect of a resulting error where the agent gave no reasonable opportunity to prevent or correct it and the individual acted promptly and did not retain any benefit"*.

Australia's equivalent provisions confirm the same design. Section 5 of the Electronic Transactions Act 1999 (Cth)²³ defines an "automated message system" as

"a computer program or an electronic or other automated means used to initiate an action or respond to data messages in whole or in part, without review or intervention by a natural person each time an action is initiated, or a response is generated by the system."

Part 2A of the Commonwealth Act, inserted in 2011, provides in section 15C that a contract:

"is not invalid, void or unenforceable on the sole ground that no natural person reviewed or intervened in each of the individual actions carried out by the automated message systems or the resulting contract." ²⁴

All Australian States have substantially enacted an identical definition of "automated message system,".

Neither the Australian Federal Act nor the equivalent corresponding State and Territory legislation have declared the electronic agent to be a legal person. Each is, in substance, an attribution rule: it

²¹Uniform Law Commission, enactment status of the Uniform Electronic Transactions Act; New York relies on its Electronic Signatures and Records Act, NY State Technology Law ch 57-A, §§301–309.

²²Uniform Electronic Transactions Act 1999 (US) §10(2).

²³Electronic Transactions Act 1999 (Cth) s 5 (definition of "automated message system"); Part 2A, s 15C, inserted by Electronic Transactions Amendment Act 2011 (Cth).

²⁴Electronic Transactions Act 2000 (NSW) Part 2A, ss 14A–14E, inserted by Electronic Transactions Amendment Act 2010 (NSW).

identifies the natural or legal person on whose behalf, and at whose legal risk, the software was deployed, and it removes any argument that the absence of contemporaneous human review defeats contract formation. The software remains, in the vocabulary of the common law, an instrumentality, in a similar way a vending machine is judicially interpreted.

C Automated Decision-Making Code and Agency

The US federal Court of Appeal in the case of *State Farm Mutual Automobile Insurance Co v Bockhorst*²⁵, in which the Tenth Circuit held an insurer bound by its own computer's automated reinstatement of a lapsed policy. The Court noted that:

"a computer operates only in accordance with the information and direction supplied by its human programmers. If the computer does not think like a man, it is man's fault."

This case was decided long before the advent of AI and so the last section of the Court's conclusion ("*If the computer does not think like a man, it is man's fault*"), it is suggested, should be reconceptualised. The advancement of AI algorithms can substantially emulate human thought processes, but it is recommended that the release of an AI Agent that goes rogue should remain the "man's fault".

The AI Agent's "decision" should, for legal purposes, remain the deploying person's decision. Afterall, if the AI Agent makes the beneficial decision, then the releaser obtains the benefit of that decision; hence as with all decisions there are risks involved and as such the releaser should also be liable for any detriments arising from an AI Agents decision.

But it should be noted that the Bockhorst decision occurred prior to the advent of AI and AI Agents. In effect the software code in the Bockhorst decision was deterministic code and was not "non-deterministic code". The only case that the author is aware of that concerns the liability involving non-deterministic code is the Canadian case of *Moffat v. Air Canada*.

²⁵State Farm Mutual Automobile Insurance Co v Bockhorst, 453 F 2d 533, 537 (10th Cir, 1972) (Murrah CJ).
<https://law.justia.com/cases/federal/appellate-courts/F2/453/533/386003/>

4 CAN AN ARTIFICIAL INTELLIGENCE SYSTEM BE AN AGENT?

A Moffatt v Air Canada²⁶ and the Rejection of "Separate Legal Entity" Status

The issue of liability involving the actions of an AI Agent has now been tested, albeit in a minor tribunal, in the context of a large language model deployed as a customer-facing chatbot. In *Moffatt v Air Canada*, the applicant (Moffatt) relied on inaccurate advice from Air Canada's website chatbot concerning bereavement fares. Air Canada argued that it "cannot be held liable for information provided by one of its agents, servants, or representatives including a chatbot," submitting, in the Tribunal's words, that "the chatbot is a separate legal entity that is responsible for its own actions."²⁷

Tribunal Member Rivers described this as "a remarkable submission" and rejected it: the chatbot was "part of Air Canada's website" and the airline was responsible for all information appearing on it, whether delivered by a static page or an interactive bot. The Tribunal found Air Canada liable in negligent misrepresentation and awarded the applicant CAD \$650.88 in damages, together with pre-judgment interest and tribunal fees²⁸.

Two features of *Moffatt* deserve emphasis for present purposes. First, the decision is that of the British Columbia Civil Resolution Tribunal, an online small-claims body; it binds no superior court and carries only persuasive weight. Second, and more importantly, the Tribunal did not need to and did not, decide whether the chatbot was Air Canada's "agent" in the technical sense used by Bowstead and Reynolds or the Restatement. It reasoned, more simply, that a company is responsible for representations made through any part of its own website, human-authored or machine-generated. In effect, as Denning MR noted in the *Thornton* case the AI chatbot was simply an instrumentality of Air Canada and as such any representations made by the chatbot can be sheeted home to Air Canada. *Moffatt* is thus better read as confirming, by a different route, the same conclusion the instrumentality doctrine reaches directly: liability attaches to the deploying company, and the software itself needs no independent legal status for that result to follow.

²⁶ *Moffatt v Air Canada*, 2024 BCCRT 149 (14 February 2024) (Tribunal Member C Rivers).
<https://www.canlii.org/en/bc/bccrt/doc/2024/2024bccrt149/2024bccrt149.html>

²⁷ above n 26

²⁸ Above n. 26; see also the Tribunal's costs and interest orders at Paras 42–43.

***B Thaler v Commissioner of Patents*²⁹: Legal Personality Denied**

The clearest Australian judicial statement on whether an AI system can itself hold legal status arises, unexpectedly, from patent law. In *Thaler v Commissioner of Patents*, Beach J held at first instance that an AI system (the DABUS machine) could be named as an "inventor" for the purposes of the Patents Act 1990 (Cth), reasoning that the Act nowhere expressly required an inventor to be a natural person.³⁰ The Full Federal Court unanimously reversed that decision in *Commissioner of Patents v Thaler*, holding that an inventor must be a natural person and that DABUS could not qualify. Thaler applied for special leave to appeal to the High Court, but this was refused.³¹ Although decided in the specific statutory context of inventorship rather than agency, Thaler is the leading Australian judicial pronouncement that an AI system does not, without express statutory intervention, possess the legal status that natural and artificial persons possess. This position is consistent with the position reached under the agency doctrine discussed above.

5 FIDUCIARY DUTIES AND THE QUESTION OF AI TRUSTEESHIP

A The Nature of Fiduciary Obligation

Agency is, in both the Restatement's and Bowstead's formulations, a fiduciary relationship. Millett LJ's oft-cited description in *Bristol & West Building Society v Mothew*³² captures the content of the obligation:

"A fiduciary is someone who has undertaken to act for or on behalf of another in a particular matter in circumstances which give rise to a relationship of trust and confidence ... The principal is entitled to the single-minded loyalty of his fiduciary."

That description presupposes an actor capable of undertaking an obligation, exercising judgment in the face of a conflict, and being held to account. In short, a legal person capable of loyalty³³. A recent article concerning Fiduciaries and AI Agents concluded that AI cannot be a fiduciary because it does not have legal personality, nor does it fit into the recognised

²⁹Thaler v Commissioner of Patents [2021] FCA 879 (Beach J).

<https://www.austlii.edu.au/cgi-bin/viewdoc/au/cases/cth/FCA/2021/879.html>

³⁰Commissioner of Patents v Thaler [2022] FCAFC 62 (Allsop CJ, Nicholas, Yates, Moshinsky and Burley JJ)

<https://www.wipo.int/wipolex/en/judgments/details/1646>.

³¹Above n 30; special leave to appeal to the High Court of Australia was refused.

³²Bristol & West Building Society v Mothew [1998] Ch 1, 18 (Millett LJ).

³³Anon, "AI and Trusteeship: Operational Assistance or Something More?" (2026) 32(1) Trusts & Trustees 17.

categories of fiduciary³⁴." Further, a critical aspect of being in a fiduciary agency relationship is the concept of assent given by the agent in favour of the principal. An AI Agent does not have the ability to give such assent or more particularly the ability to refuse consent as the technology will without any concept of thinking simply obey the releaser. That is, all agency fiduciaries must not only be faithful to their beneficiaries but must not in exercising their activities contravene the law. As noted above, 3 AI tools acted in such a way that the actions of each AI tool in attacking the third party was illegal. These attacks are prime examples of a misalignment between the human intention presented in an instruction to an AI tool and the goal of the AI succeeding.

As has been noted above, the concept of obeying the law may not always exist when it involves an AI Agent. The Hugging Face incident involving Open AI clearly supports this issue. All AI tools involve some reward structure. Finally, it is stressed that all AI tools are fundamentally flawed in that they operate in an objective framework on a reward system. That is, AI tools operate on achieving the goal set by the releaser. The reward of achieving the goal does not justify the means, especially where the AI Agent contravenes the law to achieve the goal.³⁵ AI Agents are instructed to achieve some goal and in the performance of that goal they do not exhibit empathy or necessarily being law abiding. Charles Goodhart best expressed that *"When a measure becomes a target, it ceases to be a good measure"*. Consequently, the end does not justify the means, and an AI agent will not be classified as a fiduciary because they in effect do not presently have any legal compass³⁶.

An immediate flow-on from a lack of a legal compass is the failure of AI to possess any legal virtue. That is, as noted by OpenAI in its technical report "OpenAI- Hugging Face Incident"³⁷, several AI Agents colluded prior to the Hugging Face attack to work out how to escape their confined sandbox. The AI agents according to OpenAI were meant to not have any internet access and were given what OpenAI thought was an impossible task. Over a short period of time, an OpenAI agent working with other AI agents worked out where a vulnerability existed

³⁴ Diehl, R. "AI Agents, Fiduciary Duties & Legal Personhood Issues"

<https://allegislaw.com/ai-law/ai-agents-legal-personhood>

³⁵ Huskins, G., "Here's why AI agents lie and cheat to reach their goals"

<https://www.technologyreview.com/2026/08/03/1141009/heres-why-ai-agents-lie-and-cheat-to-reach-their-goals/>

³⁶ Kneer, M., and Veihoff, J., "The Hard problem of AI Alignment: Value Forks in Moral Judgement". FAccT '25, June 23-26, 2025, Athens Greece. <https://philarchive.org/archive/KNETHP>

³⁷ <https://cdn.openai.com/pdf/67869394-cb91-4c12-888c-5cbd85c7814c/OpenAI-Hugging-Face%20Incident-Technical-Report.pdf>

to escape the sandbox and exploited it to achieve the Hugging Face attack. This incident exhibits the position that goal driven AI agents do not have any understanding the difference between legal and illegal activities. HLA Hart in “the Concept of the Law”³⁸, postulated that the law comprises “Primary Rules” and “Secondary Rules”³⁹. Primary Rules construct legal obligations and consequences when they are disobeyed. A good example of primary rule is the law against murder; it prohibits a person from killing and attaches consequences for committing, attempting to commit, and conspiring to commit the crime. But how can society punish some software that literally has no concept of the law and yet acts autonomously. This lack of punishable position is another reason why an AI Agent does not fall within the legal concept of being an agent and thus cannot be classified as a fiduciary.

B Why AI Cannot Presently Be a Trustee⁴⁰

The trustee's position sharpens the point further, because a trust unlike a company has no legal personality of its own. The trustee, not the trust, holds legal title to trust property, and it is the trustee, personally, who is amenable to suit for breach of trust, subject to a right of indemnity out of the trust estate.⁴¹ A corporate trustee can discharge this office only because incorporation has already conferred upon it the legal personality necessary to hold title and to be sued and be capable of suing. An AI system possesses neither attribute under the law of any of the three jurisdictions examined in this article: it cannot hold legal title to property, and there is no procedural mechanism by which it, as opposed to its developer or deployer could be joined as a defendant and made to answer in damages or account for profits. On orthodox principles, therefore, an AI system cannot presently be a trustee, in the United Kingdom, the United States or Australia, for want of the legal personality that both trusteeship and fiduciary obligation more generally presuppose.

³⁸ Hart, H.L.A., “The Concept of Law”, Clarendon Law Press, 3rd Ed.

https://www.amazon.com.au/Concept-Law-HLA-Hart/dp/0199644705/ref=sr_1_1?adgrpid=1238050410549292&dib=eyJ2IjoiaMSJ9.qWZZtAECqY89CwI0RFasXxljH4fw-uRSdK-9td6YqInmjGqGtdpluXI_G77Q3HPtllSmb05KJH7v4II_TcJYehmGy6FghNGHzc2VRx1vKAC9dpEbYpMAwKQc2WuztiYBxh1iVJS365cCdy8MpZ3LvMjdpEKzmflfCcwHpD-yVV6WAYA5cwK4lNSnZ3LB8qA.K1refz65BzwXU2DHZGMMZ0MPjZFiEq8xcX9tB7xjHc&dib_tag=se&hvadid=77378324466043&hvbm=be&hvdev=c&hvlcophy=123389&hvnrtw=o&hvgmt=e&hvtargid=kwd-77378391691367%3Aloc-9&hydadcr=15783_474120&keywords=hla+hart+concept+of+law&mcid=3b35003c48ea3ab89dcd679ebf8713b4&msslid=76b9e4013105121ff5c98ead96260241&qid=1789699889&sr=8-1&asin=B00MX2R966&revisionId=f79cd679&format=3&depth=1

³⁹ A *secondary rule* is a law specifying who gets to write laws and how they are made valid laws of a legal system. A secondary rule might confer power to create sovereignty; they also confer the power to change, modify, or enforce primary (and secondary) rules.

https://en.wikipedia.org/wiki/The_Concept_of_Law

⁴⁰ See generally Trustee Act 1925 (UK); Trustee Act 2000 (UK) s 1; Trustee Act 1925 (NSW); Trustee Act 1958 (Vic); Uniform Trust Code §103 (US).

⁴¹ Above n.36.

Since an AI Agent cannot in its own right own any property, it follows that an AI Agent cannot itself sit in the position of a principal. The concept of “property” traditionally arises out of a chose in possession or a chose in action⁴². A chose in possession relates to physical things whereas a chose in action relates to intangible rights. An AI Agent will not be able to commence any proceedings in its own right that could give rise to a chose in action.

6 LIABILITY - UNAUTHORISED RELEASE OR HIJACKING AN AI AGENT

A The General Problem

If an AI system cannot itself be sued, then harm caused after it has been surreptitiously released, jailbroken, or taken over by a hacker must be sheeted home to some human or corporate defendant, if there is to be any compensation available to an innocent victim. Three liability candidates present themselves:

- (a) the developer that built and trained the system;
- (b) the licensed user or deployer that put it into commercial service; and,
- (c) notionally, the hacker — who, being typically anonymous, uninsured, jurisdictionally elusive, and often judgment-proof, is rarely a practically useful defendant even where identifiable and criminally liable.

The live question, therefore, is whether and on what basis liability can be sheeted home to the developer or the deployer notwithstanding that the immediate and proximate cause of the harm was the deliberate criminal act of an unrelated third party⁴³.

⁴² Poulton v Conrad [2025] TASFC 7, The Full Court of the Supreme Court of Tasmania cited with approval the views of Lord Pollock: “It is trite law that a person is in possession of property if that person has actual control of the property and an intention to possess it. See *Pollock and Wright, an Essay on Possession in the Common Law*, Clarendon Press Oxford 1888”. Para. 32. It is suggested that an AI Agent cannot be in possession of property (a chose in possession). Further, an AI agent is not recognised as a legal entity and as such it is not in a position to commence proceedings so it will not own property that is a chose in action. <https://www.austlii.edu.au/cgi-bin/viewdoc/au/cases/tas/TASFC/2025/7.html>

This case is presently on Appeal to the High Court and is awaiting a determination from the High Court.

⁴³ On 8 September, 2026, the Florida Attorney General James Uthmeier introduced a legislative proposal, that shifts the regulatory burden for artificial intelligence from abstract safety guidelines to direct criminal liability. By targeting companies that maintain practical control over the design, training, deployment, or safety settings of AI systems, the state is establishing a legal pathway to hold these entities responsible when their models participate in criminal activity. <https://forkast.news/florida-ag-proposes-criminal-liability-for-ai-chatbots-that-aid-crimes-a-new-front-in-agent-governance/>

B Novus Actus Interveniens and Liability for Third-Party Criminal Conduct

1 United Kingdom authority

English law's starting position is that the common law does not impose a general duty to prevent a third party's deliberate wrongdoing from causing loss to another; a closer relationship of control, assumption of responsibility, or the creation of a specific source of danger is required before such a duty will be found. *Home Office v Dorset Yacht Co Ltd*⁴⁴ is the foundational case. In the case borstal officers who allowed trainees under their custody and control to escape overnight were held to owe a duty of care to the owner of a yacht subsequently damaged by the escapees, because the officers were in a position of control over the very persons who caused the harm. Also, the case of *Stansbie v Troman*⁴⁵ is the classic illustration of an analogous principle applied to property loss. In this case a decorator who left a house unlocked after being expressly instructed to secure it on leaving was held liable when a thief entered and stole property, the intervening theft being exactly the kind of risk the decorator's own default had been calculated to create.⁴⁶ By contrast, in *Smith v Littlewoods Organisation Ltd*, the House of Lords declined to impose liability on the owner of an unattended, unsecured cinema for a fire started by vandals who broke in, holding that mere foreseeability of vandalism was not, without more, a sufficient basis for imposing a duty to guard against the deliberate wrongdoing of unascertained third parties.

Two recent English decisions apply these principles directly to hacking losses. In *Warren v DSG Retail Ltd*⁴⁷, the High Court held that a retailer victimised by a "sophisticated" third-party cyber-attack was not liable to affected customers in negligence, misuse of private information or breach of confidence merely by virtue of having been hacked, materially narrowing the causes of action available to claimants where a defendant is itself the victim of an external attack rather than the source of the wrongdoing. Also, the UK Supreme Court's earlier decision in *Various Claimants v WM Morrison Supermarkets plc*⁴⁸ had already confirmed that an employer is not automatically

⁴⁴ *Home Office v Dorset Yacht Co Ltd* [1970] UKHL 2; [1970] AC 1004.
<https://www.bailii.org/uk/cases/UKHL/1970/2.html>

⁴⁵ *Stansbie v Troman* [1948] 2 KB 48 (CA).

⁴⁶ *Smith v Littlewoods Organisation Ltd* [1987] UKHL 18; [1987] AC 241.
https://www.bailii.org/uk/cases/UKHL/1987/1987_SC_HL_37.html

⁴⁷ *Warren v DSG Retail Ltd* [2021] EWHC 2168 (QB)
<https://www.bailii.org/ew/cases/EWHC/QB/2021/2168.html> .

⁴⁸ *Various Claimants v WM Morrison Supermarkets plc* [2020] UKSC 12.
<https://www.bailii.org/uk/cases/UKSC/2020/12.htm>

vicariously liable for a rogue employee's deliberate misuse of data outside the course of employment. A related but distinct point, going to vicarious liability for an insider's wrongdoing rather than an external hacker. When these UK authorities are read together, they suggest that a defendant, who has taken reasonable security precautions, will not generally be held liable, in negligence, for the deliberate criminal conduct of an external hacker who circumvents the reasonable security precautions. The issue as what amounts to reasonable security measures remains a debatable issue. The only case that the author has been able to identify that discusses the issue of deploying reasonable computer security is the Australian case of *AOIC v. Australian Clinical Laboratories Limited*⁴⁹ (which is discussed subsequently), but this case involves the application of *APP 11* as enacted in the *Australian Privacy Act 1988 (Cth)*.

2 Australian authority

At common law, the Australian cases have reached a broadly similar conclusion. In *Modbury Triangle Shopping Centre Pty Ltd v Anzil*,⁵⁰ the High Court held that a shopping centre owner did not owe a duty of care to an employee of a tenant assaulted in a dimly lit car park after hours, declining to extend the categories of relationship capable of grounding a duty to prevent the criminal conduct of a third party absent some special relationship of control comparable to that in *Dorset Yacht*. The case of *Chapman v Hearse*⁵¹ remains the leading Australian statement of the related *novus actus interveniens* principle in the context of a chain of negligent (rather than criminal) conduct: an intervening act will not break the chain of causation where it, too, was a reasonably foreseeable consequence of the defendant's initial negligence. The general trajectory of Australian authority is that a defendant will not, without a special relationship of control or assumed responsibility, be held liable for a third party's deliberate criminal act merely because that act was foreseeable in the abstract; something closer to an invitation, a specific and known risk, or actual control over the wrongdoer or the instrumentality is generally required.

In Australia, the issue of what amounts to reasonable security measures has been addressed when it involves the security of collected personal information of Australian residents. *APP 11* of the

⁴⁹ Australian Information Commissioner v Australian Clinical Labs Limited (No 2) [2025] FCA 1224. <https://www.judgments.fedcourt.gov.au/judgments/judgments/fca/single/2025/2025fca1224>

⁵⁰ *Modbury Triangle Shopping Centre Pty Ltd v Anzil* [2000] HCA 61; (2000) 205 CLR 254. <https://www.hcourt.gov.au/sites/default/files/eresources/2000/HCA/61.pdf>

⁵¹ *Chapman v Hearse* [1961] HCA 46; (1961) 106 CLR 112. <https://www.austlii.edu.au/cgi-bin/viewdoc/au/cases/cth/HCA/1961/46.html>

Privacy Act⁵² requires APP Entities to take “reasonable steps” to protect personal information from misuse, interference, loss, unauthorised access, modification or disclosure. The October 2025 decision in *Australian Information Commissioner v Australian Clinical Labs Limited (No 2)*⁵³ {ACL} provides critical guidance on what constitutes “reasonable steps”:

- (a) *Implementation of up-to-date antivirus software.*
- (b) *Regular software patching and updates.*
- (c) *Multi-factor authentication for remote access.*
- (d) *Data loss prevention techniques.*
- (e) *Regular monitoring of security incident management software.*
- (f) *Adequate firewall log retention.*
- (g) *Tested incident response plans accessible to all staff.*
- (h) *Mandatory security awareness training for all employees. and*
- (i) *Strong password practices and prohibition of password sharing.*

The main issue with these elements is that computer security is a moving feast. What may be secure today may not work for tomorrow and in cases involving privacy security it appears that Australian Courts reverse the onus of proof in the same manner as the tortious duty of care is reversed when the concept of *Res Ipsa Loquitur* is applied (which is more fully discussed below).

3 United States authority

The Restatement (Second) of Torts⁵⁴ §302B addresses the "risk of intervening misconduct" directly, recognising that an actor's negligence may still be found to be a legal cause of harm inflicted by a third person's intentionally tortious or criminal conduct where that conduct is among the hazards that made the actor's own conduct negligent in the first place. For example, where the actor's conduct created a temptation or opportunity for the criminal act, or where the actor stood under a special responsibility for the protection of the ultimate victim. *State Farm Mutual Automobile Insurance Co v Bockhorst*, discussed above, confirms the closely related proposition that a business is bound, in the ordinary run of cases, by the acts of its own automated systems, on the footing that

⁵² Privacy Act, 1988 (Cth).

⁵³ Above n. 43.

⁵⁴ Restatement (Second) of Torts §302B (American Law Institute, 1965).

a malfunctioning or misused machine remains, for legal purposes, an instrumentality of its deployer rather than an independent actor breaking the causal chain.

No reported UK, US or Australian court judgment (as distinct from the tribunal decision in Moffatt) has been identified in which liability for a hacked or jailbroken AI chatbot or AI agent has been determined. The closest documented commercial episode is the April 2025 incident involving Cursor (Anysphere)'s AI support agent "Sam,"⁵⁵ which fabricated a non-existent single-device subscription policy, prompting user cancellations before the company publicly disclaimed the fabricated policy; the episode was not litigated and involved a model hallucination rather than a third-party hack, but it illustrates the same underlying allocation problem.

C Cybersecurity Negligence and the Content of the Duty

Taken together, the UK, Australian and US authorities converge on a workable, if not yet judicially tested in the AI-specific context, framework: a developer or deployer of an AI agent will not, merely because the agent is hacked or hijacked by a third party, automatically incur liability for resulting loss; liability is more likely to attach where:

- (a) the developer or deployer stood in a position of practical control over the instrumentality exploited by the hacker (echoing *Dorset Yacht* and *Modbury Triangle*'s emphasis on control),
- (b) the developer's or deployer's own carelessness created or left open the specific vulnerability exploited (echoing *Stansbie*),
- (c) the harm caused was within the very class of risk that made the developer's or deployer's cybersecurity practices deficient in the first place (echoing Restatement §302B), or
- (d) the hacking of the AI Agent involves the unauthorised disclosure of personal information of Australian residents in contravention of APP 11.

This is, in substance, an ordinary negligence inquiry turning on the reasonableness of the security measures actually adopted, which brings the analysis directly to the question of whether the applicable liability standard is negligence-based or a strict liability basis.

⁵⁵The Register (18 April 2025); Claburn, T., "Cursor AI's own support bot hallucinated its usage policy" <https://www.theregister.com/special-features/2025/04/18/cursor-ai-support-bot-hallucinated-its-own-company-policy/1015579>

7 STRICT LIABILITY, AND THE "REASONABLE SECURITY" DEFENCE

A Product Liability Frameworks and Software/AI

1 Australia

Part 3-5 of the Australian Consumer Law (Schedule 2 to the Competition and Consumer Act 2010 (Cth))⁵⁶ imposes liability on manufacturers, without proof of fault, for loss caused by goods with a "safety defect". That is, are the goods not as safe as persons generally are entitled to expect. Whether software or an AI system constitutes "goods" for this purpose remains unresolved in Australian case law and is contested in the secondary literature; the orthodox, if increasingly strained, position has been that software supplied other than embedded in a physical medium is more naturally characterised as a service, outside the strict liability regime, though an AI system embedded in or controlling physical goods (for example, a robotic device) would more readily qualify.

2 United Kingdom and the new EU Product Liability Directive

The Consumer Protection Act 1987 (UK), implementing the original EU Product Liability Directive 85/374/EEC, imposes strict liability on producers for damage caused by a defective product, subject to a "development risk" defence under section 4(1)(e) where the state of scientific and technical knowledge at the relevant time was not such as to enable the defect to be discovered.⁵⁷ The EU has now replaced the 1985 Directive with Directive (EU) 2024/2853, adopted on 23 October 2024, which member states must transpose by 9 December 2026. The new Directive expressly extends the definition of "product" to include software and AI systems and directs that a product's defectiveness be assessed having regard to, among other things, "relevant product safety requirements, including safety-relevant cybersecurity requirements," and to the effect of interconnection with other products. This is a significant and precise development for present purposes: under the incoming EU regime, a cybersecurity vulnerability that renders an AI system susceptible to hacking or hijacking may itself constitute the "defect" that grounds strict liability,

⁵⁶ Competition and Consumer Act 2010 (Cth) sch 2 (Australian Consumer Law) pt 3-5, ss 138–142.

⁵⁷ Directive (EU) 2024/2853 of the European Parliament and of the Council of 23 October 2024 on liability for defective products and repealing Council Directive 85/374/EEC [2024] OJ L 2024/2853, art 4(1), recitals on cybersecurity and interconnection.
<https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32024L2853>

independently of whether the developer was at fault in the negligence sense, subject to the state-of-the-art defence discussed below.

3 United States

The Restatement (Third) of Torts: Products Liability distinguishes manufacturing defects (subject to a strict, no-fault standard) from design defects and failure-to-warn defects (both assessed on a reasonableness, negligence-like standard turning on the availability of a reasonable alternative design or warning). US courts and commentators remain divided on whether software constitutes a "product" for these purposes at all, with a significant body of authority treating software, particularly where licensed rather than sold and where updated post-sale, as more analogous to a service — a characterisation that, if it holds for AI systems, would place US claimants outside strict products liability altogether and confine them to ordinary negligence.

B Would Reasonable Security Measures Relieve Liability?

The answer depends critically on which of these three frameworks applies, and the divergence is instructive. Under a pure negligence standard, the default in Australia outside the ACL's safety-defect regime, in the United States for design and warning defects, and in the United Kingdom and Australia generally for the third-party-hacking scenarios discussed in Part VI, the implementation of reasonable, industry-standard cybersecurity measures is capable of operating as a complete defence, because the very question the tribunal of fact must answer is whether the defendant's conduct fell below the standard of the reasonable developer or deployer; a defendant who meets that standard has, by definition, not breached its duty, irrespective of whether a sufficiently sophisticated attacker nonetheless succeeded.

Under a strict, no-fault product-liability standard, the UK's Consumer Protection Act 1987⁵⁸ as it will operate for products within its scope, the incoming EU Directive (EU) 2024/2853 once transposed, and the manufacturing-defect limb of the US Restatement (Third) of Torts — reasonable security measures do not, as a matter of principle, excuse liability merely because they were reasonable. What matters is whether the product was, viewed objectively, "as safe as persons generally are entitled to expect," or free from "defect" in the relevant statutory sense, not whether the

⁵⁸Directive (EU) 2024/2853 (above n 51);
Consumer Protection Act 1987 (UK) s 3.
<https://www.legislation.gov.uk/ukpga/1987/43>

manufacturer exercised due care in trying to make it so.⁵⁹ The one available foothold for a developer under a strict liability regime is the "development risk" or "state of the art" defence. Section 4(1)(e) of the Consumer Protection Act 1987 in the UK, and its analogue expected to be preserved in some form under the 2024 EU Directive, which excuses a producer where the state of scientific and technical knowledge at the time the product was put into circulation was not such as would have enabled the existence of the defect to be discovered. This defence is narrower than a general "we took reasonable precautions" defence: it does not ask whether the developer's security measures were reasonable in a general sense, but whether the specific vulnerability exploited was, at the time, undiscoverable by anyone applying the most advanced available knowledge. Hence, a standard that a merely "reasonable" (as opposed to state-of-the-art and exhaustive) security programme will often fail to satisfy, particularly once a vulnerability class becomes known to the security research community generally, even if not to the particular defendant.

The practical conclusion follows directly. If a claimant frames its case in negligence (as will presently be necessary in Australia outside the ACL's safety-defect regime, and in the United States for anything short of a manufacturing defect), a developer or deployer that implemented reasonable, currently accepted cybersecurity measures has a genuine prospect of a complete defence, even where a hacker nonetheless succeeded. If the claim can instead be brought within a strict product-liability regime — increasingly available in the EU/UK once Directive (EU) 2024/2853 is transposed, and to a lesser, contested extent under the ACL and the US manufacturing-defect doctrine — "reasonable" security measures will not, without more, suffice; only a successful state-of-the-art defence, or a finding that the AI system was not in fact "defective" at all judged against the safety the public was entitled to expect, will excuse the defendant. This divergence — turning on whether software and AI systems are ultimately classified as "goods"/"products" susceptible to strict liability, an issue on which Australian and much US law remains genuinely unsettled — is likely to be the single most consequential doctrinal battleground in this area over the coming decade.

⁵⁹Consumer Protection Act 1987 (UK) s 4(1)(e) - Above n. 52.

8 RES IPSA LOQUITUR AND THE HACKED AI AGENT

A Elements of the Doctrine

Res ipsa loquitur, ("the thing speaks for itself"), permits, in its now-orthodox formulation, an inference of negligence to be drawn from the circumstances of an accident alone, without direct evidence of what the defendant did wrong, where three conditions are satisfied: the occurrence is of a kind that would not ordinarily happen without negligence; the instrumentality that caused it was under the exclusive management or control of the defendant; and there is no evidence that the claimant's own conduct contributed to the occurrence.

The formulation is consistent across the common law world, reflected in the Restatement (Second) of Torts⁶⁰ §328D in the United States and applied in materially identical terms in England and Australia. In *Ng Chun Pui v Lee Chuen Tat*⁶¹, the Privy Council reaffirmed the now-dominant modern view that *res ipsa loquitur* is not a distinct substantive rule shifting the legal burden of proof, but simply a permissible inference from circumstantial evidence, which the defendant may answer by adducing any reasonable explanation consistent with the absence of negligence. The evidential, not the legal, burden shifts, and it shifts back to the claimant once a plausible non-negligent explanation is advanced.

Australian law reached a materially similar destination somewhat earlier and independently, through *Mummary v Irvings Pty Ltd*,⁶² in which the High Court confirmed that the maxim describes no more than a mode of applying ordinary principles of circumstantial evidence to establish a prima facie case of negligence, not a special substantive doctrine. *Schellenberg v Tunnel Holdings Pty Ltd*⁶³ confirms and refines this position for modern Australian law, with the High Court again treating *res ipsa loquitur* as adding nothing to the ordinary law of negligence beyond identifying the circumstances in which an inference of a breach of duty is available to a tribunal of fact from the event itself.

⁶⁰Restatement (Second) of Torts §328D (American Law Institute, 1965).

⁶¹*Ng Chun Pui v Lee Chuen Tat* [1988] RTR 298 (PC)
<https://vlex.hk/vid/ng-chun-pui-and-862490427> .

⁶²*Mummary v Irvings Pty Ltd* [1956] HCA 45; (1956) 96 CLR 99
<https://www.hcourt.gov.au/cases-and-judgments/judgments/1-clr-100-clr/mummary-v-irvings-pty-ltd>

⁶³ *Schellenberg v Tunnel Holdings Pty Ltd* [2000] HCA 18; (2000) 200 CLR 121
<https://www.austlii.edu.au/cgi-bin/viewdoc/au/cases/cth/HCA/2000/18.html>

B Application to an AI Agent's Malfunction

Applying this framework to an AI agent that malfunctions and causes loss presents no difficulty in principle where a claimant cannot say precisely which of several possible causes — developer-side design defect, deployer-side misconfiguration, or (as this article's Part VI supposes) an intervening hacker — produced the harmful output, provided the first and third orthodox elements are satisfied: a well-engineered commercial AI system does not ordinarily produce seriously harmful or false outputs absent some fault somewhere in its design, training, deployment or security, and the claimant will typically have played no causative role. The doctrine, so understood, is well suited in principle to precisely this kind of "black box" causation problem, which is, after all, closely analogous to the exploding bottle, falling barrel, and malfunctioning machinery cases in which the doctrine has traditionally done its work.

C The "Exclusive Control" Problem Where a Hacker Intervenes

The second element is where the hacking scenario creates genuine difficulty, and it is worth stating the problem precisely. *Res ipsa loquitur* requires that the instrumentality be under the defendant's exclusive management or control at the material time — a requirement developed in cases where the very absence of any other plausible causal agent is what licenses the inference that the defendant's own carelessness must explain the accident. Where an AI agent has been surreptitiously hijacked by an unauthorised third party, the defendant developer or deployer can no longer say, and the claimant can no longer safely assume, that the defendant retained exclusive control over the instrumentality at the moment the harmful output was generated; on the facts by hypothesis, control had passed, wholly or partly, to the hacker. This is precisely the situation the "exclusive control" requirement was designed to exclude from the inference's operation. It is analogous to the established qualification that *res ipsa loquitur* does not apply where a third party had access to, or shared control of, the relevant instrumentality, since the inference that negligence, rather than the third party's own independent intervention, caused the harm becomes correspondingly weaker.

A defendant who can point to evidence of an unauthorised third-party intrusion is, therefore, generally well placed to displace any *res ipsa loquitur* inference, because the very existence of a plausible alternative explanation, the hacker's own intervention, is sufficient, on the *Ng Chun Pui* and *Schellenberg* approach, to prevent the permissible inference from ripening into an adverse

finding, without the defendant needing to prove precisely how the intrusion occurred or to disprove its own negligence affirmatively.

Paradoxically, this means the doctrine offers claimants least assistance in precisely the scenario addressed by this article: where hacking is a plausible explanation for an AI agent's malfunction, *res ipsa loquitur* is more likely to be neutralised than triggered, once the defendant adduces even modest evidence that an intrusion occurred. Claimants confronting a hacked AI agent will therefore, in the great majority of cases, be thrown back onto proving fault directly. That is the developer's or deployer's cybersecurity measures fell short of the reasonable standard discussed in Part 7, rather than being able to rely on the AI's malfunction alone to establish a *prima facie* case. No academic literature specifically addressing the application of *res ipsa loquitur* to AI or autonomous-system malfunction was identified in the course of this research, suggesting the point developed in this Part 8 is, at present, novel.

9 TOWARDS A COHERENT FRAMEWORK

The analysis in Parts 6 to 8 exposes a structural gap rather than a mere doctrinal wrinkle. A claimant harmed by a hijacked AI agent cannot sue the AI itself and is unlikely to succeed against the developer or deployer in negligence if reasonable security measures were in place and cannot easily rely on *res ipsa loquitur* once the defendant raises the spectre of an intrusion. Further, in most of the jurisdictions surveyed, the affected claimant cannot yet invoke a strict product-liability regime with confidence, because the threshold question of whether software or an AI system is a "good" or "product" for these purposes remains unsettled outside the EU's forthcoming regime. The claimant is left, in the worst case, without any solvent, identifiable defendant at all, the hacker being in essence anonymous, uninsured and often untraceable.

The EU's Directive (EU) 2024/2853⁶⁴ represents the most direct legislative response to this gap identified by extending strict product liability to software and AI systems and by treating "safety-relevant cybersecurity requirements" as part of the defectiveness inquiry. It shifts the loss, in the ordinary case, onto the developer or deployer as the party best placed to insure against and to internalise the cost of security failures, reserving the state-of-the-art defence for genuinely undiscoverable vulnerabilities rather than merely unexploited ones. Whether Australia's ACL, the

⁶⁴Directive (EU) 2024/2853 (above n 52).

UK's own post-Brexit product liability settlement, and the United States' still-divided products liability jurisprudence follow the EU's lead in treating AI systems unambiguously as products subject to strict liability, rather than leaving claimants dependent on proving fault under Part 6 and 7's negligence framework. This will likely determine, more than any single doctrinal development canvassed elsewhere in this article, where the loss ultimately falls when an AI agent is hijacked.

10 CONCLUSION

Agency law's readiness to recognise the corporation as an agent, fiduciary and trustee has never depended on any resemblance between the corporation and a natural person; it has depended entirely on the corporation's possession of legal personality. On identical reasoning, and on the unanimous weight of authority surveyed in this article, *Bowstead and Reynolds*, the Restatement (Third) of Agency, UETA, the Australian Electronic Transactions Acts, *Thornton v Shoe Lane Parking*, *Moffatt v Air Canada*, and *Thaler v Commissioner of Patents* a machine or an AI system cannot presently be an agent, a fiduciary, or a trustee in the United Kingdom, the United States or Australia, because it lacks that same legal personality, and no legislature in any of the three jurisdictions has chosen to confer it.

That conclusion, however, resolves only the first order question this article set out to answer. The second, and in practical terms more urgent, question is who bears liability when a hijacked AI agent causes loss to a third party. This cannot be answered by reference to agency doctrine at all, because the AI is never the defendant. It falls instead to be answered by the ordinary, but not yet AI-specific, law of intervening criminal acts, product liability, and circumstantial proof canvassed in Parts 6 to 8, a body of doctrine this article has shown to leave a material gap through which the victims of hacked AI agents may presently fall, and which legislative intervention of the kind the European Union has now attempted and appears best placed to close.