An Autonomy-Based Classification: AI Agents, Liability and learnings from the UK Automated Vehicles Act

Julia Smakman*
Ada Lovelace Institute
jsmakman@adalovelaceinstitute.org

Lisa Soder*†
Interface
lsoder@interface-eu.org

Connor Dunlop*
Ada Lovelace Institute
cdunlop@adalovelaceinstitute.org

Weiwei Pan
Harvard
weiweipan@g.harvard.edu

Siddharth Swaroop
Harvard
siddharth@seas.harvard.edu

Noam Kolt
Hebrew University
noam.kolt@mail.huji.ac.il

Abstract

The prospect of AI agents—autonomous systems capable of independently executing complex open-ended tasks with only limited human involvement—presents significant challenges for liability law. As AI agents become more capable, exhibit greater autonomy and act in increasingly complex environments, new questions arise regarding the attribution of liability if and when these systems cause harm. Drawing on existing governance regimes for autonomous vehicles, we propose an "autonomy scale" for AI agents to help structure legal discussions on allocation of liability. Specifically, we analyse how key concepts in tort liability may apply to AI agents and use the UK's Automated Vehicles Act 2024 to illustrate how an autonomy scale could be implemented in practice. This preliminary analysis, we suggest, is a useful first step in developing legal categories of AI agents and establishing appropriate liability frameworks.

1 Introduction

AI agents—autonomous systems capable of independently executing complex, open-ended tasks with only limited human involvement—have attracted growing interest and investment in research [25, 19], industry [14, 31, 33] and policy [40, 10, 7]. Early examples of AI agents, such as Cognition's Devin, MultiOn's Agent Q, and Sakana's AI Scientist exhibit some, albeit limited degree of autonomy, enabling them to independently perform a variety of activities in software engineering, online retail, and scientific research. The economic benefits of AI agents will increase as the technology improves and agents become capable of performing a wider range of tasks more reliably.

At the same time, the delegation of tasks to AI agents and the reduction in human oversight introduce significant risks and uncertainties [15], not least for our legal system. At the moment,

*Lead authors with equal contribution. Order of first three authors randomised. Authors are free to list themselves first in the author order in their CVs.

†Correspondence should be directed to lsoder@interface-eu.org

2nd Workshop on Regulatable ML at NeurIPS 2024.

in the absence of legislation or case law, it is unclear who will bear the liability for the actions of AI agents among the agent's user, deployer, or developer. Against this backdrop, we examine the challenges that AI agents pose for tort liability.

One area where lawmakers have grappled with comparable policy challenges is liability for autonomous vehicles (AVs). In particular, we seek to draw lessons from the UK's Automated Vehicle Act 2024 [14]. Through the Automated Vehicle Act 2024, the 'user-in-charge' will not be held (criminally) liable for damages caused by the AV when it is in self-driving mode. Instead, the manufacturer or software developer are directly liable for offences resulting from 'the way the vehicle drives'. An accompanying insurance scheme through the Automated and Electric Vehicles Act 2018 also protects AV users against civil liability claims. This legal framework thus establishes that when a 'user-in-charge' of an AV has no effective control over how the vehicle operates, they should be protected from liability.

We argue that a similar risk-based approach that focuses on autonomy as a measure of control—as implemented in AV laws—can inform how we might assess liability for AI agents. Risk-based approaches are frequently used by regulators to guide their interventions [3], as seen in areas like medical devices, finance, environmental regulation, and more recently, AI [5, 23]. For instance, the EU's proposed AI Liability Directive aims to align liability with the risk-based tiers established by the AI Act [13]. Current AI legislation typically considers the inherent properties of AI systems, such as their intended use case or capabilities (operationalized via compute thresholds). However, more advanced agents, specifically when built upon general-purpose AI systems, might introduce challenges around the degree of user control that these lenses might not account for. An autonomy-based classification might better address these challenges by shifting the focus from the AI system's inherent capabilities to the level of control afforded to the user in specific deployment contexts, as advocated by Morris et. al [28].

This paper proceeds as follows. Section 2 provides a definition of key concepts related to AI agents and tort liability. Section 3 explores the primary challenges that autonomous AI agents present for existing liability frameworks. Section 4 analyzes the UK's regulatory approach to AV and its implications for liability. Section 5 proposes a taxonomy of autonomy for AI agents, drawing inspiration from AV classifications and discusses the merits and limitations of such an approach. Finally, Section 6 summarizes our findings and offers recommendations for future research in this rapidly evolving field.

2 Key Definitions

2.1 AI agents

AI agents can be broadly defined[3] as systems that can independently plan and carry out a sequence of actions on behalf of users, often without necessitating continuous human supervision. They are characterised by their ability to perceive and operate in complex environments across a variety of domains, to adapt their strategies and actions based on new input autonomously, and to interact with their surroundings, for instance, through natural language interfaces. [38, 20, 39, 15].

However, defining what exactly constitutes an agent is often a more complex question in practice. Previous work [7, 22, 39] has argued that defining an agent cannot be tied to a binary attribute. Rather, 'agenticness', and consequently the autonomy of a system, can be defined by an interplay of various characteristics, such as

- *Goal underspecification:* The ability to operate based on high-level, underspecified goals without detailed instructions. This includes functioning on open-ended tasks in the absence of constant human supervision. [11, 6]
- *Action Complexity:* The scope and potential impact of actions the system can perform, encompassing tool use (e.g., web search, programming) and operation across varied environments [15, 22, 9]
- *Adaptability:* in their approach to pursuing a goal, by being not only able to make decisions that are "temporally dependent upon one another" [6], but also capable of behaving differently when circumstances change.

[3] For a more detailed, interdisciplinary discussion on the definition of agents, see Chopra & White p 5-27 [8]

2.2 Liability

"Liability" refers to legal responsibility one has for their actions, often requiring a party to remedy a harm it has committed or contributed to. Liability can arise from various sources, including criminal law, contract law, and tort law. For purposes of this piece, we focus on tort law as a branch of civil law that allows those harmed by negligent or dangerous activities to seek compensation and redress without a contractual relationship.

Focus on negligence. Within tort law, multiple forms of liability exist [44, 32][4]. Most relevant here, the tort of negligence creates 'a general duty to take reasonable care to avoid causing harm to other people's person or property' [27][5] . This means that even in the absence of specific legislation imposing liability on certain actors for certain activities, actors can be liable under negligence if they failed to take 'reasonable case' and caused harm to someone's person or property. Reasonable care (generally) is defined by the level of caution a reasonably prudent person would exercise to prevent harms that are a foreseeable consequence of their conduct.

Establishing a "standard of care". To assign liability within tort law, it is thus important to establish the standard of care, which is the care that is taken by a reasonable actor in that position. This standard of care can be informed by a range of input: from industry standards and [35], academic research (for example on Human-Computer Interaction [34, 26]), statements by policymakers, and legal requirements [9]. Normally, a standard of care will emerge over time and evolve as we learn more about a new activity or product and the kinds of risks associated with it.

3 What Challenges Do (Autonomous) AI Agents Introduce for Liability?

The promise of more autonomous and capable AI agents introduces a host of complex challenges, such as lowering barriers for malicious activities and increasing propensity for systemic risks[6, 24, 9].[6] Against the backdrop of assigning liability, this section focuses on two particular challenges, namely the complexity of the value chain and principal-agent problems.

Information Asymmetry Similar to traditional principal-agent scenarios, AI agents often possess information that their principals (users or operators) do not. This information asymmetry complicates liability as it can be difficult to determine the agent's 'knowledge' or what it 'should know.' [2, 24].

Complex value chains The "many hands" problem—where responsibility is distributed among multiple actors [29, 24]—is not unique to AI agents. However, the intricacy of value chains [4] involved in the design, deployment, and use of AI agents exacerbates this issue. When liability is assigned, downstream deployers—often the last party in the value chain—may face a disproportionate share of responsibility. This imbalance could arise from power disparities in contractual negotiations [9, 42, 4]. Smaller entities, which typically serve as deployers, are often less equipped to bear such liabilities compared to larger, resource-rich AI providers..

Delegation and Discretion. AI agents' ability to interpret ambiguous instructions in achieving their goals is a critical feature but also a significant source of risk. Misinterpretation of user intentions can lead to actions that deviate from the intended outcomes, further complicating the assignment of liability.

[4]Both common law and civil law systems have 'theories of liability', developed through a combination of jurisprudence (decisions by judges) and statutes (laws). Some common theories of liability include fault liability (which includes intentional torts or negligent torts), strict liability (liability not requiring 'fault', usually for dangerous activities or goods), product liability (liability of manufacturers for defective products), and vicarious liability (liability for conduct of others). See Appendix

[5]Fault-based liability also creates heightened duties of care where the actor has a special relationship with the harmed party or has specialized professional training

[6]We recognize there is a wealth of literature from legal scholars exploring various forms of liability to address the challenges raised by autonomous systems, AI, and AI agents. For a good overview of such literature see footnote 5 of [24] and [8]

4 Liability & Autonomy: A Case-study in Automated Vehicles

As explained in Section 2, one of the defining features of AI agents is that they can act autonomously in complex settings. An area that has similarly grappled with the governance of an autonomous system with varying capacity for human oversight is the governance of AVs.

Regulations concerning liability for AVs offer a useful, though imperfect, analogy for AI Agents: AVs' range of actions is more limited (e.g. breaking, steering, accelerating) and their area of deployment more clearly defined (roads and driveways). Also, AVs operate in the physical world, and AI Agents may - at least at first - be predominantly inhabiting online environments. Still, AVs operating at Level 3 autonomy and higher, like AI Agents, have considerable autonomy in how they reach a goal set for them, do not (necessarily) get human approval for actions taken in the course of reaching that goal, and can autonomously adapt and respond to their environment. Additionally, both AVs and AI Agents grapple with when control should be kicked back to the driver/operator, and under what circumstances it is safe to let them chart out their own path.

A key tool developed to structure policy and liability discussions for AVs is the concept of levels of autonomy, which this section will further explore. Autonomous Vehicles ('AVs') are classified in various ways based on autonomy. The Society of Automotive Engineers (SAE) splits AVs into six levels of autonomy, whilst the Association of British Insurers differentiate 3 levels. The full SAE and ABI taxonomies with descriptions can be found in the appendix1 2.

4.1 AV Liability frameworks in the UK based on levels of Autonomy.

Automated Vehicle Act 2024 [17] ('the Act') [7] Focused on criminal liability, this Act focuses on self-driving cars and does not cover 'assisted driving' (roughly comparable to Level 0, 1, 2 in the SAE taxonomy). Within AV use, the Act distinguishes between the 'user-in-charge' ('UiC') mode and the 'no-user-in-charge' ('NUiC') mode. Chapter 10 of the Act explains that "an individual is the 'user-in-charge' of a vehicle if: (1) the vehicle is an authorised AV with an authorised user-in-charge feature, (2) that feature is engaged, and (3) the individual is in, and in position to exercise control of, the vehicle, but not controlling it." Thus, the user-in-charge does not control the driving and does not actively need to monitor the driving, but - when clearly signalled by the AV with sufficient time - should be able to take over control. This is roughly comparable with AVs that the ABI would classify as 'automated', and the SAE as Level 4 and 5 (potentially, as Level 3 AVs, if they do not require the driver to actively monitor the driving).

The Act establishes, amongst others, that as the UiC is not in control of the 'steering, accelerating, or breaking', they cannot be held criminally liable (i.e. be prosecuted) for any offences committed when the car is in self-driving mode [17] [30].[8] Instead, the 'Authorised Self-Driving Entity', which can be the vehicle manufacturer or software developer or a partnership between the two, is responsible for 'the way the vehicle drives' and offences resulting from that [18].

The Automated and Electric Vehicles Act 2018 The Automated and Electric Vehicles Act 2018 concerns civil liability (such as liability under tort law, i.e. the possibility of being sued by another private party for damages) and thereby complements the 2024 Act [16]. The 2018 Act requires AVs to be insured to quickly settle claims for damages caused by AVs. The injured party has a direct claim against the insurer, so the insurer pay-out is immediate. The insurer 'may then bring a secondary claim against any person or body responsible for the incident' to recover costs, such as the vehicle manufacturer [30]. To support this, the 2024 Act (Section 14) imposes requirements on the sharing of information (for example about vehicle safety) with public authorities and private businesses like insurers. This duty to share information can help insurers determine which party in the chain is liable [18].

[7] Although the Act is pre-emptive (AVs that can drive autonomously at SAE Level 4 and 5 are not driving on UK roads yet) and was passed very recently, analysis of its effectiveness thus not yet possible. Still, the Act was the result of an extensive multi-year consultation with various stakeholders and it was welcomed as a framework for clarifying legal and safety norms and thereby enabling the introduction of AVs on British roads.

[8] Although the UiC has no responsibility regarding the 'manner of driving', they are not without responsibility. The explanatory notes accompanying the Act clarify that the UiC is still responsible for 'insuring the vehicle, checking that any load is secure before they set off, and ensuring that any children in the vehicle are wearing seatbelts' (Explanatory Notes, par. 32).

Assignment of liability along the supply chain is grounded in levels of autonomy. In short, the UK government has recognized that it is undesirable to place liability, criminal or civil, with the 'driver' of an AV when the AV has a high level of autonomy and the person in the car has no control over how the AV is driving. Instead, a system is proposed where criminal liability falls on the vehicle and/or software developer. With regards to civil liability (e.g. tort liability), a compulsory insurance scheme has been instigated that makes the insurer the first to pay out (making damages easy to recover for any affected party), whilst the insurer is afterwards able to seek recourse against the vehicle and software developers. This is supported by information sharing obligations that the 2024 Act imposes on manufacturers and/or software developers.

5 Taxonomy: An Autonomy-Based Classification of AI Agents

In this section, we argue that it is useful to think about how AI agents might be categorised based on autonomy levels. Like with AVs, the level of autonomy of an AI agent may determine to what extent an end-user might be able to exercise human oversight[9]. This taxonomy for AI agents can serve as input to allow for more nuanced future discussions about the allocation of liability along the supply chain. As exemplified by the UK's Automated Vehicle Act 2024, the level of autonomy and user control can and should play an important role in deciding what forms of liability should attach to different products.

Why focus on autonomy as a proxy for risk? Current efforts to regulate general-purpose AI systems, such as the EU AI Act, the proposed AI Liability Directive, and the US Executive Order, adopt a risk-based approach. This involves using proxies for risk, such as a model's capability levels (e.g., measured by the amount of training compute), to determine the appropriate level of regulatory obligations. while valuable for understanding the potential of an AI system, might overlook how these capabilities are applied in real-world contexts—specifically, how users interact with the system and the degree of control they retain over its actions. Capabilities essentially enable certain actions but do not inherently determine the risk associated with these actions; this risk is contingent upon the deployment environment and the specific human-AI interaction model [28, 26]. Key factors such as interface design and the level of autonomy granted to the AI system significantly influence how its capabilities manifest in practical applications and, consequently, affect the level of risk posed.

Given these considerations, we focus on autonomy—defined here as the degree of independence an AI system possesses in performing tasks—as our primary lens for discussing liability. While more difficult to operationalize (e.g., capabilities are currently proxied via compute used in a training run); this approach might allow to account for both the system's capabilities and the real-world context in which it operates.

A risk-based framework to structure liability discussion. Combining these levels of autonomy with the allocation of liability in automated vehicles law 4 provides some rules of thumb in assigning liability across the AI agent value chain. In the UK's Automated Vehicle Act 2024, users are liable when they are 'in charge' of the vehicle. Applied to AI agents, we might similarly expect liability to accrue with users who are more clearly 'in charge' of the agent's actions (level 1-2 agents) but to be distributed away from the user when they have less control over the agent's actions (level 3 agents). However, this rule of thumb is overly simplistic: there are clearly situations when a user should be liable for harms caused by a level 3 agent (e.g. choosing to use a level 3 agent in an area with foreseeable harms, or with overly open-ended goals)

Need for further research to establish liability within the framework. Active research is therefore needed to go beyond rules of thumb and towards an established duty of care for AI agent control. In this regard, we suggest some avenues, based on how control is determined in AV law. For example:

- How can we define that a user is 'in charge', or that they are in a position to 'exercise control'?
- In AV law, users are not liable when they cannot control 'steering, accelerating, or breaking'. Can we define similar parameters to determine when a user is in control for AI agents?
- Some levels of AV alert a human driver to take over control in certain situations: what should the equivalent be for AI agents?
- Some levels of AV, and their 'user-in-charge' features need to be 'authorised' by the regulator. Should level 3 autonomy need authorisation for use in certain high-risk (or open ended) contexts?

	Definition	Example	Who has control?	Level of control?
Level 3: Fully autonomous	An AI agent that directly impacts the world with minimal human intervention (especially when enacting complex goals in complex environments)	Tasked with "maximising return on investment" on a retail web platform over a period of time, an AI independently conducts complex research, interfaces with manufacturers, negotiates contracts, and markets a product. [9]	Users have control over what tasks the agent is used for, and what its high-level goals should be, and (possibly) approval for critical decisions. How the AI agent acts is informed by how the developer trained the model (e.g. capabilities), and how the application provider fine-tuned and made available the system (e.g. user control mechanisms)	A key determiner of control is the ability to monitor, intervene, and approve important actions. Level 3 agents operate in open-ended and complex environments (e.g. interacting with humans) with reduced ability for human intervention and course correction than for level 2 agents.
Level 2: Conditionally autonomous	An AI agent that takes direct actions and accomplishes specified goals under human supervision. Humans retain significant oversight and can intervene.	An AI agent with computer use capabilities analyses market data, searches the web, and purchases a product as directed by a human user. Human approval is needed for the purchase, but the agent may take dozens of steps to complete the task, accessing a range of tools and websites to do so. [10]	Design choices by the developer and application provider mean that the user has to approve the final decision for a narrowly specified goal. The user can also monitor and override actions by the agent. The human user usually has control, but ability this may depend on many factors (e.g.interface design). Malfunctions would likely be outside the user's control.	The goal is much more specified than for a level 3 agent (e.g. purchase a specific product).The actions the agent takes will be complex at times (e.g. navigating and interacting with websites) but the agent will not be able to solve problems beyond computer use (e.g. interacting with human customer service operators). The narrowly specified goal needs approval by the human user.
Level 1: Decision support for narrowly defined tasks	An AI assistant that performs specific functions that are narrowly defined by human operators, primarily offering support for well-defined, short-term tasks without independent decision-making.	An AI trading assistant analyses financial information provided by the user (e.g. stock prices), flagging potential trading opportunities based on pre-set criteria. It can execute buy or sell orders when explicitly instructed by a human trader, but cannot make any independent trading decisions.	The agent can take some narrowly defined actions, but these are single-or-few step, and clearly designated and approved by the human operator.The user therefore remains fully in control and responsible for any decisions affecting external parties.	Human user takes information provided by an AI system as input and makes the final decision on any Actions. The goal is clearly specified, and the assistant cannot achieve complex tasks or adapt to new circumstances.

Table 1: Levels of AI Autonomy

When seeking to answer these questions, developers and application providers should always use the test of how a 'reasonable average person' would interact with the AI system. Such research can help define a standard of care for human control and subsequently allocation of liability for AI agent.

5.1 Merits of an autonomy-based taxonomy for AI Agents

Establishing a standard of care for each level of agent autonomy. The standard of care for the development and deployment of an agent that can take few-step actions to achieve a narrow task (level 1) should be different than that for an agent that can pursue open-ended goals in complex environments (level 3), as in the latter the impacts are much less predictable, and redress is likely to be more difficult. This framework gives an avenue for establishing a more nuanced and targeted standard of care per each level of autonomy, thus informing the allocation of legal liability.

Incentivizing technical work on control mechanisms for AI agents. If an autonomy framework were used to inform law, developers might be incentivised to work on controllability for agents (in order to reduce exposure to liability). This could shape industry best practices: for example, it may be the case that meaningful control by the end-user will not be possible for level 3 agents, and this could push developers and deployers in the direction of providing level 2 agents. If all levels of agents had the same standard of care in the law, innovation for controllability may be less likely to happen.

5.2 Limitations

AV framework applicability. We recognise that our analysis for AVs does not map perfectly onto all AI agents: AVs pose an obvious and direct risk to life and, although they also operate in a complex environment requiring complex decisions, they operate in a somewhat bounded domain. However, we think it provides a useful framework to explore trade-offs in AI autonomy and human control that can inform liability.

Operationalizing autonomy levels. Further, a key challenge for using such a framework to investigate liability is operationalizing the different levels of agent autonomy [28]. Developing ecologically valid benchmarks for agent autonomy remains an open research question [37, 36], as it requires consideration of not only capabilities across a wide range of tasks, but also agent affordances (e.g., tools, deployment constraints) [22] and human-AI interaction paradigms [21, 43].

6 Conclusion

Increasing autonomy of AI agents pose challenges for tort law: complexity of the value chain and principal-agent problems make it difficult to locate parties responsible for agent-involved harms. As this is an emerging technology, the standard of care that actors in the value chain should exercise when developing or operating AI agents still lacks clarity, creating further uncertainty about what we may legally expect from such actors. Inspired by automated vehicle taxonomies and legislation, we provide an autonomy taxonomy for AI agents that can help further the development of the standard of care for different actors in the AI agent value chain. Mirrored in the UK's AV legislation, we support a shifting away (although not complete elimination) of liability from the end-user, when the end-user is not capable of exercising effective control over the actions of the AI agent. We provide some suggestions for further research.

A Types of liability

Type of liability	Components	Control	Example
Fault-based (negligence)	1. Standard of care based on reasonable person standard 2. Breach of duty (fault: intent or negligence) 3. Damage 4. Causation	Duty of care to prevent harms that are reasonably foreseeable, i.e. tortfeasor could and should have known that the harm could materialise and should have taken reasonable precautions. It was within his control to prevent the harm from happening and failed to do so.	Cafe owner leaves open a cellar hatch whilst restocking and a customer accidentally falls into it and injures themselves; owner should have foreseen this created a risk and taken reasonable precaution (close the hatch or put a warning sign).
Strict liability	1. Duty of care attached to object or activity (through case law or statute) 2. Damage 3. Causation	Control plays a less obvious role: liability is assigned based on law or statute (regardless of fault/reasonable precautions taken), usually for a dangerous object or activity.	Dog bites a person, owner took all reasonable precautions, but is still held liable if victim proves injury and that this has been caused by the dog. An accident happens at a chemical plant causing physical injury through chemical exposure in surrounding villages. Despite the plant having taken all necessary precautions and adhering to industry safety standards, it is automatically held liable for all physical injury damages caused by the chemical exposure.
Vicarious liability (agency law)	1. Wrongful act committed by agent that caused foreseeable damage 2. Within scope of agency 3. Principal had the ability to control the agent	The principal needs to have effective control over the conduct of the agent, meaning that he could (and should) have the ability to meaningfully impact how the agent conducts their work.	An electrician wires something in a faulty way and causes a fire, the company they work for is held liable by the homeowners for property damage.
Product liability	1. The product is defective 2. The defect caused the damage 3. The defect was present when the product left the manufacturer's control	The manufacturer is liable for defects that occurred when the product was within his control.	A portable charger catches on fire during normal use and causes damage. The manufacturer is held liable (unless the manufacturer can prove the charger was not defective when it left the manufacturer's control).

Table 2: Types of Liability

B Taxonomy of Automated Vehicles

Both the Society of Automotive Engineers (SAE) and Association of British Insurers (ABI) have established taxonomies for levels of driving automation for self-driving cars. These taxonomies can be found in Figure 1 and Figure 2 below.

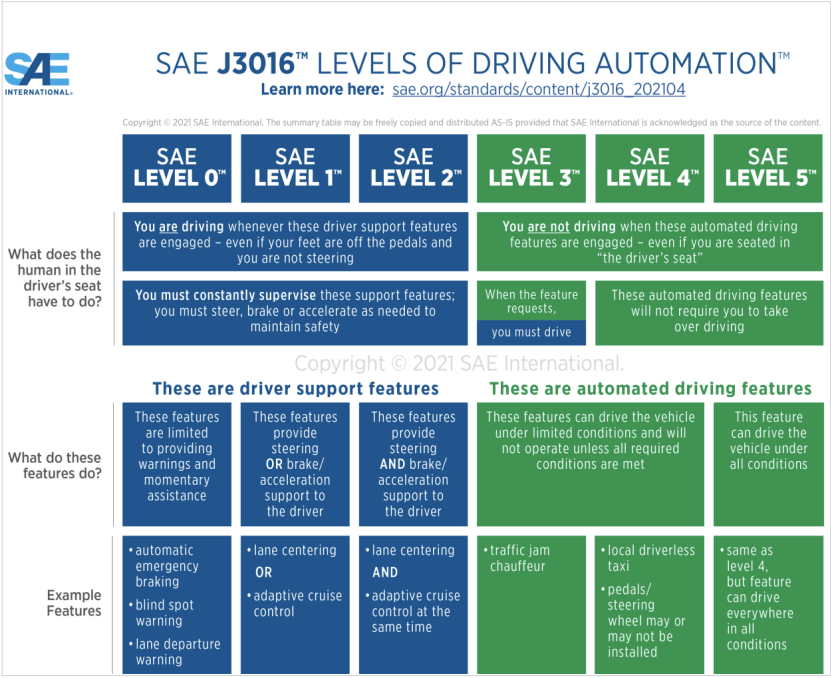

Figure 1: SAE Levels of Driving Automation

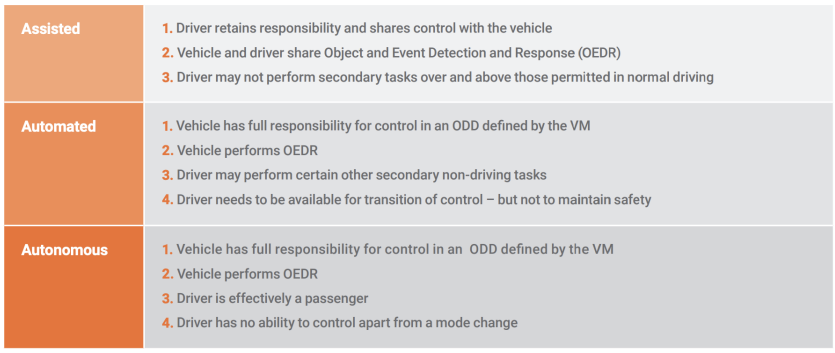

Figure 2: ABI Levels of Driving Automation

References

[1] Anthropic. Introducing computer use, a new claude 3.5 sonnet, and claude 3.5 haiku, October 2024. URL `https://www.anthropic.com/news/3-5-models-and-computer-use`. Accessed: 2024-11-02.

[2] Kenneth Joseph Arrow. The economics of agency. *(No Title)*, 1984.

[3] Julia Black. The emergence of risk-based regulation and the new public risk management in the united kingdom. *Public law*, page 512, 2005.

[4] Ian Brown. Allocating accountability in ai supply chains, 2023.

[5] Johanna Chamberlain. The risk-based approach of the european union's proposed artificial intelligence regulation: Some comments from a tort law perspective. *European Journal of Risk Regulation*, 14(1):1–13, 2023.

[6] Alan Chan, Rebecca Salganik, Alva Markelius, Chris Pang, Nitarshan Rajkumar, Dmitrii Krasheninnikov, Lauro Langosco, Zhonghao He, Yawen Duan, Micah Carroll, et al. Harms

from increasingly agentic algorithmic systems. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pages 651–666, 2023.

[7] Alan Chan, Carson Ezell, Max Kaufmann, Kevin Wei, Lewis Hammond, Herbie Bradley, Emma Bluemke, Nitarshan Rajkumar, David Krueger, Noam Kolt, et al. Visibility into ai agents. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 958–973, 2024.

[8] Samir Chopra and Laurence F White. *A legal theory for autonomous artificial agents*. University of Michigan Press, 2011.

[9] Peter Cihon. Chilling autonomy: Policy enforcement for human oversight of ai agents. In *41st International Conference on Machine Learning, Workshop on Generative AI and Law*, 2024.

[10] Michael K Cohen, Noam Kolt, Yoshua Bengio, Gillian K Hadfield, and Stuart Russell. Regulating advanced artificial agents. *Science*, 384(6691):36–38, 2024.

[11] Alexander D'Amour, Katherine Heller, Dan Moldovan, Ben Adlam, Babak Alipanahi, Alex Beutel, Christina Chen, Jonathan Deaton, Jacob Eisenstein, Matthew D. Hoffman, Farhad Hormozdiari, Neil Houlsby, Shaobo Hou, Ghassen Jerfel, Alan Karthikesalingam, Mario Lucic, Yian Ma, Cory McLean, Diana Mincu, Akinori Mitani, Andrea Montanari, Zachary Nado, Vivek Natarajan, Christopher Nielson, Thomas F. Osborne, Rajiv Raman, Kim Ramasamy, Rory Sayres, Jessica Schrouff, Martin Seneviratne, Shannon Sequeira, Harini Suresh, Victor Veitch, Max Vladymyrov, Xuezhi Wang, Kellie Webster, Steve Yadlowsky, Taedong Yun, Xiaohua Zhai, and D. Sculley. Underspecification presents challenges for credibility in modern machine learning. *arXiv preprint arXiv:2011.03395*, 2020. doi: 10.48550/arXiv.2011.03395. URL `https://arxiv.org/abs/2011.03395`.

[12] Wes Davis. Google is reportedly developing a 'computer-using agent' ai system, October 2024. URL `https://www.theverge.com/2024/10/26/24280431/google-project-jarvis-ai-system-computer-using-agent`. Accessed: 2024-11-02.

[13] European Commission. Liability rules for artificial intelligence, 2024. URL `https://commission.europa.eu/business-economy-euro/doing-business-eu/contract-rules/digital-contracts/liability-rules-artificial-intelligence_en`. Accessed: 2024-11-03.

[14] Hayden Field. Ai agents are having a 'ChatGPT moment' as investors look for what's next after chatbots. *CNBC*, jun 2024. URL `https://www.cnbc.com/2024/06/07/after-chatgpt-and-the-rise-of-chatbots-investors-pour-into-ai-agents.html`. Accessed: September 9, 2024.

[15] Iason Gabriel, Arianna Manzini, Geoff Keeling, Lisa Anne Hendricks, Verena Rieser, Hasan Iqbal, Nenad Tomašev, Ira Ktena, Zachary Kenton, Mikel Rodriguez, et al. The ethics of advanced ai assistants. *arXiv preprint arXiv:2404.16244*, 2024.

[16] James Goudkamp. Automated vehicle liability and ai. *The Cambridge Handbook of Private Law and Artificial Intelligence*, 2022.

[17] UK Government. Automated vehicles act 2024, 2024. URL `https://www.legislation.gov.uk/ukpga/2024/10`. Accessed: September 10, 2024.

[18] UK Government. Explanatory notes: Automated vehicle act 2024, 2024. URL `https://www.legislation.gov.uk/ukpga/2024/10/pdfs/ukpgaen_20240010_en.pdf`. Accessed: September 10, 2024.

[19] Ziniu Hu, Ahmet Iscen, Aashi Jain, Thomas Kipf, Yisong Yue, David A Ross, Cordelia Schmid, and Alireza Fathi. SceneCraft: An LLM agent for synthesizing 3D scenes as blender code. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp, editors, *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 19252–19282. PMLR, 21–27 Jul 2024. URL `https://proceedings.mlr.press/v235/hu24g.html`.

[20] Qiuyuan Huang, Naoki Wake, Bidipta Sarkar, Zane Durante, Ran Gong, Rohan Taori, Yusuke Noda, Demetri Terzopoulos, Noboru Kuno, Ade Famoti, et al. Position paper: Agent ai towards a holistic intelligence. *arXiv preprint arXiv:2403.00833*, 2024.

[21] Lujain Ibrahim, Saffron Huang, Lama Ahmad, and Markus Anderljung. Beyond static ai evaluations: advancing human interaction evaluations for llm harms and risks. *arXiv preprint arXiv:2405.10632*, May 2024. URL `http://arxiv.org/abs/2405.10632`.

[22] Sayash Kapoor, Benedikt Stroebl, Zachary S Siegel, Nitya Nadgir, and Arvind Narayanan. Ai agents that matter. *arXiv preprint arXiv:2407.01502*, 2024.

[23] Leonie Koessler, Jonas Schuett, and Markus Anderljung. Risk thresholds for frontier ai, 2024. URL `https://arxiv.org/abs/2406.14713`.

[24] Noam Kolt. Governing ai agents. *Available at SSRN*, 2024.

[25] Jessy Lin, Yuqing Du, Olivia Watkins, Danijar Hafner, Pieter Abbeel, Dan Klein, and Anca Dragan. Learning to model the world with language. In *Forty-first International Conference on Machine Learning*.

[26] Arianna Manzini, Geoff Keeling, Lize Alberts, Shannon Vallor, Meredith Ringel Morris, and Iason Gabriel. The code that binds us: Navigating the appropriateness of human-ai assistant relationships. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 7, pages 943–957, 2024.

[27] Markus Anderljung Matthew van der Merwe, Ketan Ramakrishnan. Tort law and frontier ai governance, 2024.

[28] Meredith Ringel Morris, Jascha Sohl-Dickstein, Noah Fiedel, Tris Warkentin, Allan Dafoe, Aleksandra Faust, Clement Farabet, and Shane Legg. Position: Levels of AGI for operationalizing progress on the path to AGI. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp, editors, *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 36308–36321. PMLR, 21–27 Jul 2024. URL `https://proceedings.mlr.press/v235/morris24b.html`.

[29] Helen Nissenbaum. Accountability in a computerized society. *Science and engineering ethics*, 2:25–42, 1996.

[30] Law Commission of England and Scottish Law Commission Wales. Automated vehicles: Joint report.

[31] OpenAI. Research into Agentic AI Systems, dec 2023. URL `https://openai.smapply.org/prog/agentic-ai-research-grants/`. Accessed: September 6, 2024.

[32] Ariel Porat and Alex Stein. *Tort liability under uncertainty*. Oxford University Press, USA, 2001.

[33] Maria Abi Raad, Arun Ahuja, Catarina Barros, Frederic Besse, Andrew Bolt, Adrian Bolton, Bethanie Brownfield, Gavin Buttimore, Max Cant, Sarah Chakera, et al. Scaling instructable agents across many simulated worlds. *arXiv preprint arXiv:2404.10179*, 2024.

[34] Kaspar Raats, Vaike Fors, and Sarah Pink. Understanding trust in automated vehicles. In *Proceedings of the 31st Australian Conference on Human-Computer-Interaction*, pages 352–358, 2019.

[35] Ketan Ramakrishnan, Gregory Smith, and Conor Downey. U.s. tort liability for large-scale artificial intelligence damages: A primer for developers and policymakers, aug 2024. Research Published.

[36] Anka Reuel, Ben Bucknall, Stephen Casper, Tim Fist, Lisa Soder, Onni Aarne, Lewis Hammond, Lujain Ibrahim, Alan Chan, Peter Wills, et al. Open problems in technical ai governance. *arXiv preprint arXiv:2407.14981*, 2024.

[37] Anka Reuel, Lisa Soder, Ben Bucknall, and Trond Arne Undheim. Position paper: Technical research and talent is needed for effective ai governance. *arXiv preprint arXiv:2406.06987*, 2024.

[38] Stuart J Russell and Peter Norvig. *Artificial intelligence: a modern approach*. Pearson, 2016.

[39] Yonadav Shavit, Sandhini Agarwal, Miles Brundage, Steven Adler, Cullen O'Keefe, Rosie Campbell, Teddy Lee, Pamela Mishkin, Tyna Eloundou, Alan Hickey, et al. Practices for governing agentic ai systems. *Research Paper, OpenAI, December*, 2023.

[40] Julia Smakman. Ai assistants: Helpful or full of hype? Ada Lovelace Institute Blog, aug 2024. URL `https://www.adalovelaceinstitute.org/blog/ai-assistants/`. Accessed: September 9, 2024.

[41] Mustafa Suleyman. My new turing test would see if ai can make $1 million. *Technology Review*, jul 2023. URL `https://www.technologyreview.com/2023/07/14/1076296/mustafa-suleyman-my-new-turing-test-would-see-if-ai-can-make-1-million/`.

[42] British Columbia Civil Resolution Tribunal. Moffat v air canada, feb 2024.

[43] Laura Weidinger, Joslyn Barnhart, Jenny Brennan, Christina Butterfield, Susie Young, Will Hawkins, Lisa Anne Hendricks, Ramona Comanescu, Oscar Chang, and Mikel Rodriguez. Holistic safety and responsibility evaluations of advanced ai models. *arXiv preprint arXiv:2404.14068*, 2024.

[44] Christiane Wendehorst. Ai liability in europe, 2022. URL `https://www.adalovelaceinstitute.org/resource/ai-liability-in-europe/`.