Jack Davidson

Artificial Intelligence and Delegation: Cost Benefit Internalization

Past work at the intersection of artificial intelligence (AI) and agency law has sought to understand how AI can be fit into existing systems of agency law, and whether existing rules effectively address the challenges and changes that AI creates.[1] By applying existing rules, these analyses can identify many of the pointed problems that AI creates for systems of agency, contract, and tort law, but generally ask the wrong questions if aiming to find solutions. Instead of analogizing and applying agency law's *rules*, we should analogize the *goals* of agency law to the problems created by AI to begin the process of identifying legal solutions to new technological challenges.

Cost Benefit Internalization Theory

Before applying the goals of agency law to the problem of artificial intelligence, we need to understand what those goals are. Paula Dalley argues that agency law is supported by a "cost benefit internalization theory," which shifts the foreseeable costs and benefits of an agency relationship from the agent to the principal:[2]

> The premise of the cost-benefit internalization theory is that agency law has a single underlying policy goal: to mitigate the effects of using an agent. Agency is a technology that enables increased or more productive activity, but the use of that technology has certain effects. Without technology, an individual's activities and their consequences are

[1] See, e.g., Pınar Çağlayan Aksoy, *AI as Agents: Agency Law*, *in* THE CAMBRIDGE HANDBOOK OF ARTIFICIAL INTELLIGENCE: GLOBAL PERSPECTIVES ON LAW AND ETHICS 146–160 (Larry A. DiMatteo, Cristina Poncibò, & Michel Cannarsa eds., 2022). https://law.nus.edu.sg/wp-content/uploads/2021/10/TRAIL-WPS-2102.pdf

[2] See generally Paula J. Dalley, *A Theory of Agency Law*, 72 U. PITT L. REV. 495 (2011).

> limited by her personal abilities; the use of an agent enhances a person's ability to act in the world… Without the restoration of the status quo provided by agency law, the use of agents would be either pointless or extremely expensive: principals would not have any assurance that they would receive the benefits of their agent's actions, and third parties would be unwilling to deal with anyone other than someone who was provably a principal.[3]

In effect, Dalley describes agency law as doing two things: (1) creating a default rule shifting internal and external costs of agency to the principal to enable socially optimal decisions about whether to employ agency; and (2) providing supporting fiduciary rules to minimize the costs of agency for all parties.

First, agency law forces principals to internalize the costs of agency.[4] For directly administered tasks, various branches of law including tort and contract attempt to force entities to internalize all costs of their actions (both standard costs and externalities).[5] In doing so, they attempt to encourage entities to make socially optimal decisions, assuming that there are no beneficial externalities.[6] Agency complicates the picture of the directly administered task by potentially changing both internal and external costs. For internal costs of agency, the principal generally has to pay the agent to execute the task instead of taking on the opportunity cost of spending time that could be dedicated to other tasks. For external costs of agency, the actions of the agent may increase or decrease the relative likelihood and magnitude of some externality. In Dalley's argument, agency law attempts to centralize the incremental costs of using agency in

[3] Paula J. Dalley, *A Theory of Agency Law*, 72 U. PITT L. REV. 495, 497 (2011).
[4] The law seems to focus more on costs than benefits, perhaps because it assumes that benefit internalization will occur in many cases through contracting with beneficiaries (at least as long as benefits are not diffuse).
[5] Contract law seems to focus primarily on internal costs and tort law primarily on externalities.
[6] Beneficial externalities may be better addressed via social programs or subsidies to support such activity.

one actor, the principal, and allows the principal to decide whether to delegate a task to an agent. If law is functioning well, the principal will choose to complete a task if it is socially beneficial (either under agency or direct administration) and will choose to delegate the task to an agent if the marginal social benefits of doing so outweigh the marginal social costs.[7] For Dalley and agency law, reasonable foreseeability is the main limiting factor on what costs can be shifted on to the principal. So long as something is reasonably foreseeable because of choosing to use the technology of agency, the principal is subject to both the liabilities and benefits that might come from it because foreseeable costs and benefits can be considered as part of an *ex ante* cost-benefit calculation.

Second, agency law attempts to prevent incentive and information asymmetry problems from making agency a less useful technology by introducing fiduciary duties and corresponding liability rules. In a world where the principal and agent's incentives are perfectly aligned, an agent might be more effective at a task than the principal (more profitable). However, due to incentive misalignment problems such as moral hazard and adverse selection or information acquisition and monitoring costs, the realized gains from using an agent might be less than direct administration. For example, a lazy agent subject to moral hazard might complete the task in a sloppy way, thus realizing less gain than expected for the principal.[8] Agency law attempts to align the incentives of the agent with those of the principal (largely through fiduciary duties the agent owes to the principal and resulting liabilities). In doing so, agency law forces agents to internalize the costs of their own actions that would otherwise fall on the principal due to our

[7] This is similar to Ronald Coase's theory that the boundaries of the firm are set by relative transaction costs for internal or contractual fulfillment of needs. In the same way, agency decisions should be set by relative transaction costs for direct administration versus agency. *See* R. H. Coase, *The Nature of the Firm*, 4(16) ECONOMICA 386 (1937).

[8] Agency law corrects for this problem through the agent's Duty of Care and Skill and associated liability to the principal. See Restatement (Third) of Agency § 379.

default rule of cost centralization. Because the agents are effectively the "cheapest cost avoiders" for the harms these types of externalities cause to society, agency law shifts liability for these costs onto them to drive down overall social costs.[9] Agency law also seeks to minimize the costs to third parties from dealing with agents. In doing so, it takes something resembling a cheapest cost avoider approach, with concepts like *apparent authority* and *implied authority* creating incentives for the party with the lowest cost ability to reduce confusion (often the principal) to do so.[10]

Although different in meaningful ways,[11] delegation decisions about AI are subject to many of the same costs and benefits that are identified as motivators of agency law. Even relatively simple machine learning algorithms allow individuals to do tasks at a far greater scale than their personal abilities might otherwise suggest (and this is in some sense true of all technologies).[12] As a result, delegating tasks to AI can create incremental benefits and costs, and in order for the law to incentivize socially optimal delegation to AI, it needs to perform a similar cost internalization function to that of agency law. As in agency law, where some delegations might be socially efficient and others might be socially inefficient. A socially optimal legal regime would give actors incentives to use AI in those cases. There are undoubtedly other cases where the use of AI does not outweigh the social costs, and by forcing the right party to internalize or bear the risks of these costs, an ideal legal regime would disincentivize use. Although agency law's cost minimization fiduciary duties for the agent may not apply as directly to the problem of AI (due to the lack of a legal persona or assets for AI), the concepts minimizing

---

[9] This concept is like that of in Guido Calabresi's classic analysis of tort law. *See* GUIDO CALABRESI, THE COSTS OF ACCIDENTS 155 (1970).
[10] See, e.g., *Hoddeson v. Koos Bros.*, 47 N.J.Super. 224 (N.J.Super.A.D. 1957).
[11] See infra ____
[12] See, e.g., Meta's *Community Standards Enforcement Report* showing 1.8 billion pieces of content were actioned as spam in Q4 2022 (largely actioned by algorithms rather than human review). https://transparency.fb.com/data/community-standards-enforcement/

the social costs of using new technology and the fiduciary duties for the principal remain relevant as we analyze the problem of AI.

Hypothetical AI "Agency" Relationship

Imagine a chat bot that creates algorithmically generated responses to a party's questions. We might think that there are at least four relevant (potentially non-distinct) parties in interest for a potential transaction where a customer is discussing with the chat bot:

1. The Operator (O) – The party that is using B to provide some service.
2. The Programmer (P) – The party that programmed B for O.[13]
3. The Chat Bot (B) – The program itself.[14]
4. The Customer (C) – The person who is interacting with B.

The use of a chat bot might seem to provide some obvious social benefits. Relative to a human agent, B might be able to respond to C more quickly, more accurately, and at a lower cost. There are potential social costs as well (beyond just the dislocation created by the initial replacement of humans by new technology). For example, B might also provide less accurate answers than a human agent or create other harms.[15]

[13] Although it is possible that P and O are the same legal entity (e.g., Open.ai is both the programmer and operator for Chat-GPT), it is entirely possible that they are not (e.g., a company builds a customer service chat bot on a contract for another company that then operates it). An effective legal regime would need to be able to separate out the liability of these parties.

[14] This party is unlikely to be a "person" for the purposes of agency law, but this is less important to our analysis because we are focused on the goals rather than literal application of the law. See Pınar Çağlayan Aksoy, *AI as Agents: Agency Law*, *in* THE CAMBRIDGE HANDBOOK OF ARTIFICIAL INTELLIGENCE: GLOBAL PERSPECTIVES ON LAW AND ETHICS 146–160 (Larry A. DiMatteo, Cristina Poncibò, & Michel Cannarsa eds., 2022).

[15] See, e.g., https://gizmodo.com/gpt4-open-ai-chatbot-task-rabbit-chatgpt-1850227471

Direct application of agency law to the chat bot situation is not possible for several reasons. First, AI programs are not legal persons in the sense that they do not have the power to be party to Hohfeldian entitlements.[16] Second, AI do not have title to assets that can be drawn upon if they are found to be liable for specific acts.[17] Third and relatedly, it is less clear that legal liability rules and fiduciary duties will be able to influence the cost-benefit analysis and behavior of AI agents, in part because they have no assets of their own to enter into the cost-benefit calculation. Finally, AI's structure of two parties (B and P) in the place of one human agent creates more complex questions of who should be the "agent" under agency law. Despite these limitations on applying agency law *as is* to the problem of AI, we can apply the principles of cost-benefit centralization and social cost minimization to the AI delegation problem to achieve many of the same positive ends and develop effective legal solutions to the challenges of AI.

Centralizing Costs and Benefits

Much as in human agency law, we can begin by writing the law such that it centralizes both the social costs of choosing to delegate a task to an AI in the person choosing to make the delegation (in our hypothetical, O).[18] For internal costs and benefits incurred if B operates within a clear mandate to complete the delegated task, this is simple. O would have to pay P to create B

[16] This is more of a block on applying agency law as is rather than the concepts of agency law, as we could conceivably change law to allow AI to modify and be party to entitlements. Pınar Çağlayan Aksoy, *AI as Agents: Agency Law*, *in* THE CAMBRIDGE HANDBOOK OF ARTIFICIAL INTELLIGENCE: GLOBAL PERSPECTIVES ON LAW AND ETHICS 146, 147 (Larry A. DiMatteo, Cristina Poncibò, & Michel Cannarsa eds., 2022).

[17] Some have suggested compulsory insurance for AI as a way of addressing this issue, but such insurance is really a way of reducing variability in liability rather than allocating it to AI itself. Relying on insurance still raises the questions of what party should pay for the insurance and why and may not actually reduce overall costs to society of using a technology (as agency law aims to) rather than shifting and compensating them. Pınar Çağlayan Aksoy, *AI as Agents: Agency Law*, *in* THE CAMBRIDGE HANDBOOK OF ARTIFICIAL INTELLIGENCE: GLOBAL PERSPECTIVES ON LAW AND ETHICS 146, 159 n. 104 (Larry A. DiMatteo, Cristina Poncibò, & Michel Cannarsa eds., 2022).

[18] We can think of this as an application of a broader principle that law (through liability rules, taxes, subsidies, power-confirming rules, etc.) should attempt to make a decision maker exactly sensitive to the social costs and benefits of a decision.

and pay for any ongoing costs of its operation. O would also replace whatever benefits it receives because of providing the service through a human agent with comparable benefits from providing the service through B. We can also assume that P will attempt to gauge its risks and liability and pass those on to O in indemnification requirements or increased contractual cost.

In traditional agency law, this cost centralization is both extended and limited by the principle of "reasonable foreseeability."[19] Use of reasonable foreseeability is meant to focus liability shifting only on costs that those can be foreseen and considered during *ex ante* decision making, because those are the costs that drive economic efficiency.[20] It seems reasonable to apply this same standard to delegations to AI, as both AI and human agents raise the risk of taking unauthorized actions, both for the cost and the benefit of the delegator or principal. Although AI outcomes may not be as foreseeable and deterministic as other computer programing algorithms, they appear to be (at least for now) no less foreseeable than the actions of a human agent. If we believed that AI was more inherently risky, we could seek to prevent delegators like O from disclaiming risk by labeling AI tasks as inherently dangerous, but such a step seems drastic and potentially inefficiency generating given our goal of ensuring that AI delegation is selected if and only if it is socially efficient.[21]

Fiduciary Duties, Liability Rules, and Cheapest Cost Avoiders

In cases of human agency, agency law deals with the allocation of contract, tort, or fiduciary liability between the principal, agent, and third party based on what will minimize the social costs of employing agents. Although generally costs and benefits are passed on to the

[19] Supra note ___
[20] Dalley, fn. 18
[21] See *Majestic Realty Associates, Inc. v. Toti Contracting Co.*, 30 N.J. 425 (N.J. 1959).

principal for ultimate centralization and decision making, agency law attempts to prevent the agent and third party from using the fact of agency to gain undue leverage over the principal and force the principal to pay costs or forego benefits that the principal would have chosen to avoid or capture under direct administration of the task. It does this by giving each party duties to minimize any cost that that party is best positioned to minimize while allowing parties to pass on costs to the principal once minimized.

Like human agency, AI technology will be maximally useful for society if AI capture as many benefits for the principal as possible while not allowing costs to be higher than necessary. In agency law, this problem is primarily addressed by the Duty of Care and Skill[22] as well as the Duty of Loyalty.[23] How do we allocate responsibility for exploiting new opportunities that O might want to take advantage of in our hypothetical? For example, what if C is willing to offer a contract that could be beneficial to O (and to society as a whole)? Who should be liable if B fails to accept the contract? Socially, the answer would seem to be whoever could have most cheaply avoided the contract opportunity being missed. If the contract was rejected due to negligence or other failures in programming B to respond to reasonably foreseeable uses, then P should be liable to compensate O for the lost opportunity.[24] P would most easily be able to correct for foreseeable errors created by the substandard quality of the AI by investing more in programming or testing, and can thus pass those social costs on to O at a lower cost than if O attempted to perform those same functions itself. This cost shifting also reduces incentives for moral hazard for P in creating B. If the contract is rejected for other reasons, it would seem as though O should bear the lost opportunity cost itself, as O would have likely been the lowest cost

[22] See supra note ____
[23] Restatement (Third) of Agency § 387; *See also Meinhard v. Salmon*, 164 N.E. 545 (N.Y. 1928).
[24] John Villasenor makes a similar argument for applying products liability law to AI. See https://www.brookings.edu/research/products-liability-law-as-a-way-to-address-ai-harms/ and infra ___

avoider, either by providing different inputs in operating B or by choosing a different technology for task execution (direct administration or human agency). In all cases, O should be able to capture the benefits of using AI. Thus, in cases where the actions of P and C are reasonably foreseeable by O, the law should force O to assume all contracts created by the AI as a result of the delegation in order to force O to internalize potential costs and enable third parties to maintain trust in the system.

We can do a similar analysis of cost shifting for potential tort liability created by the selection of AI to complete tasks. Imagine that C is at risk of being harmed because of interaction with B. All parties could conceivably have some influence over the likelihood and potential magnitude of this harm. For example, if P does a poor job programming the AI, this may increase the risk of harm to the consumer. Agency law addresses each of these issues by extending Calabresi's idea of "cheapest cost avoiders" to allocate liability to those making the decisions of whether to delegate a task and how to complete it. Law should provide incentives for parties to assume costs to the extent that they reduce expected harm. The question of when P can be liable for harms caused by B (as opposed to our default rule of allocating all costs and benefits to O) seems to be a case where we could directly apply products liability law. Indeed, John Villasenor at the Brookings Institute has suggested as much.[25] Villasenor gives several examples of potential products liability claims for AI programs using MRI images to try to diagnose disease arguing that cases such as design defects or negligence in programming could be used to support tort liability for AI. P can influence the quality of the AI by expending more time and effort on building and testing B prior to delivery to O (as well as continued iteration after delivery). Thus, much like the agency law duty of care and skill or direct products liability in tort,[26] an effective

[25] See https://www.brookings.edu/research/products-liability-law-as-a-way-to-address-ai-harms/
[26] See, e.g., *Hoover v. Sun Oil Company*, 58 Del. 553 (Del. Sup. Ct. 1965).

system of AI law would allocate costs related to careless and negligent programming to P. In general, if products liability law seems to allow us to hold the creator responsible for harms in the software case, we should apply similar law to hold P responsible in creation of B. However, we might think that during ongoing operation of B, O is well positioned to provide some amount of ongoing oversight to ensure that B is still achieving the desired goals and to minimize and new or unforeseen harms that might arise during operation. This is similar to the corporate law duty of oversight, which requires that directors remain informed and maintain controls over a corporation, and which allows for parties to sue directors for damages if the duty is breached.[27] Similarly, AI law could place a duty with associated liability rule that AI operations maintain reasonable information on performance of and controls over AI. As under products liability law, we might also argue for liability shifting for harm caused by C's negligence in interactions with B. As Villasenor notes, "just as a purchaser of an automobile who drives it at twice the speed limit and then gets in an accident can't reasonably blame the manufacturer, a user of an AI-based system who applies it in clearly inappropriate ways will bear responsibility for resulting harms."[28] However, liability cannot be allocated to C if C uses B in a reasonably foreseeable way and is still harmed. At that point, both under tort and products liability law, O and P are better positioned to protect against through testing and improved direction, and therefore all costs of torts created by choosing to delegate a task to an AI should be allocated to O (so long as P and C act in reasonably foreseeable ways) in the same way that agency law passes foreseeable costs and benefits to the principal.

[27] *See Marchand v. Barnhill*, 212 A.3d 805 (Del. 2019); *In re Caremark Int'l Inc. Deriv. Litig.*, 698 A.2d 959, 970 (Del. Ch. 1996).
[28] https://www.brookings.edu/research/products-liability-law-as-a-way-to-address-ai-harms/

O's decision about how to represent O's relationship with B to C should influence O's liability to the extent that O obscures the nature of the operating relationship between B and O. For example, if C believes that she is interacting directly with O, but is in fact interacting with an AI (B), O should be maximally liable for harms created by that interaction on a principle like that of *apparent agency*. Because O is the only one in a position to correct C's view that she is interacting with B rather than with O directly, O must bear increased liability for harms in such a case. In contrast, if O clearly identifies to C that she is interacting with B, then some burden and discretion shifts to C in deciding whether to use the AI-powered process. One additional note here is that as in agency law, O cannot disclaim liability for a task simply by delegating it to an AI. If the task carries inherent risk (it is inherently dangerous or ultrahazardous), then resulting harm is a foreseeable consequence of O's delegation of the task to B and cannot be avoided simply be removing direct responsibility.[29]

Conclusion

As AI becomes more common as a tool for completing tasks previously delegated to humans, law will need to adapt in order to provide the same types of cost internalization and minimization rules that agency law provides for human agents. Although agency law should not directly perform this role due to the differences between AI and legal persons, the goals and social considerations of agency law can inform us about how to respond to the revolutionary technology of AI with robust legal policy. We should apply the foreseeability principle of agency law to centralize all foreseeable costs and benefits of using AI in the person making the delegation decision (the Operator), covering both tort and contract liability. At the same time, we

[29] See *Majestic Realty Associates, Inc. v. Toti Contracting Co.*, 30 N.J. 425 (N.J. 1959).

can apply concepts from products liability and fiduciary law to allocate liability for faulty programming and unforeseeable behavior to those best equipped to minimize them (the programmer and third party). By emphasizing these two parallel systems of liability allocation, we can create law for AI delegation that offers the same social benefits that agency has for human agents, while also working within the limitations created by the lack of full personhood for current AI. As we rapidly shift towards general AI where AI increasingly moves beyond the control or influence of human programmers, we may need to adjust law yet again to account for fuller personhood of general AI and to provide the right incentives for AI "agents" that can think for themselves and behave more like people than today's relatively simple AIs.