Autonomous Artificial Intelligence and Liability: A Comment on List

[Note: This is the accepted manuscript for a forthcoming piece in Philosophy & Technology.]

Abstract: Christian List argues that responsibility gaps created by viewing artificial intelligence (AI) as intentional agents are problematic enough that regulators should only permit the use of autonomous AI in high-stakes settings where AI is designed to be moral or a liability transfer agreement will fill any gaps. This work challenges List's proposed condition. A requirement for 'moral' AI is too onerous given technical challenges and other ways to check AI quality. Moreover, transfer agreements only plausibly fill responsibility gaps by applying independently-justified group responsibility attribution norms such that AI raises no unique regulatory norms.

Autonomous Artificial Intelligence and Liability: A Comment on List

Christian List offers a responsibility gap-based argument for a condition whereby "regulators should permit the use of autonomous artificial intelligence [AI] in high-stake settings only if they are engineered to function as moral … agents and/or there is some liability-transfer agreement in place" (2021: 1213). The following argues that a requirement for 'moral' AI is too onerous and liability transfer agreements cannot provide a justified "backstop" (1215) parallel to their use in group settings. Agreements only plausibly fill responsibility gaps by applying independently-justified group responsibility attribution norms such that AI faces no unique regulatory norms.

Motivating List's Condition

List's argument for his condition is primarily grounded in the need to avoid responsibility gaps. There is, in short, a need to ensure someone is accountable for harms that would normally be attributable to human action. Private law fills responsibility gaps in group agency cases by requiring certain group accountability mechanisms. List's condition is offered as a corollary.

Per List, groups and AI both exhibit distinct intentional agency. This is the "primary parallel" (1221) between them. Groups like states, corporations, or firms and truly autonomous AI meet the basic conditions for agency, namely a combination of representational (e.g., belief) and motivational (e.g., desire) states and a capacity to act on same (1219). They can take actions that we would normally view as intentional if taken by humans but which are not plausibly attributable to particular human decisions. 'The U.S.A.' can participate in strategic interactions with 'Russia' that are not solely attributable to their executives (1215-1216). Likewise, adaptive machine learning-enabled medical tools, for example, can perform actions that cannot be attributed to any human. They can provide more accurate diagnoses of a medical condition in ways developers could not have predicted by changing performance based on real-world data.

Where group and AI actions cannot be fully attributable to individuals, List suggests, there is a risk of responsibility gaps. Following Johannes Himmelreich (2019), List suggests gaps arise iff an entity (e.g., corporation, AI) performs an act that would trigger responsibility if performed by a human (e.g., spilling oil, misdiagnosing a treatment) but no one can be held fully responsible for the act (e.g., the corporation/AI cannot be held liable and individuals are not fully responsible for its actions). Individuals are only responsible for relevant acts where they play normatively significant (e.g., enacting, authorizing, or design) roles, and then only to the extent that their roles contribute to the decision to act. AI can produce harms even where operators, owners, regulators, manufacturers, etc. all act diligently; it would be unfair to hold anyone responsible for all harms (1223, 1225-1226). As in group agency cases, this seems inevitable: to have agency that is not fully attributable to humans is just what it means to have a group agent and something similar should inform how we understand 'fully autonomous' AI distinct from its users/creators.[1]

Responsibility gaps in high-stakes cases (defined "according to society's criteria" but including many military, medical, and financial cases (1228-1230)) permit "unaccountable" decision-making and can leave those harmed unable to find proper redress (1239). List argues for a legal requirement for 'moral' AI that is fit "to be held responsible" (1239) to avoid such results. Such

[1] List himself hedges on whether autonomous AI-induced gaps are "inevitable" or merely "possible" (1223-1224).

AI must possess the capacity to make normative judgments and respond to same (moral agency proper), access to relevant knowledge, and control over their acts (1227). This is, List contends, at least conceptually possible (1230). We hold groups responsible for their actions and condition their (legal) existence to ensure responsibility. Doing similarly for AI appears desirable.

Liability transfer agreements then serve a necessary backup role for avoiding gaps where moral design is impossible. Just as owners or managers of corporations are subject to strict liability for some corporation-induced harms (e.g., illnesses caused by food safety issues), List suggests, it will be appropriate to hold someone legally responsible for some major AI-induced harms for which they would otherwise not be fully morally responsible. List offers (other) consequentialist reasons to justify this position (e.g., strict developer liability would "incentivize" safe AI development) but primarily focuses on responsibility gaps and the need to fill them (12231-1232).

Critiquing List's Condition

Requiring the development of moral AI initially appears compelling but raises problems. Assume that we can agree on a conception of 'morality' that can be legally enforced consistent with public reason. Designing moral agents still may prove technically infeasible. The costs of prohibition may not then be worth it. Moral AI is currently highly theoretical. AI able to 'reason' at all remains largely conceptual. Even tools designed to address narrow concerns are rife with issues that lead some to question whether technical solutions can address them. To wit, tools designed to address particular challenges, like the use of 'corrected' data to address bias, fail to account for many related problems, as cases of systematically-biased results from medical AI trained on corrected data make clear (e.g., Vyas et al. 2020, McCradden et al. 2020).

AI responsibility parallel to, but distinct from group responsibility, may also prove impossible. For instance, List suggests that corporations must be able to hold funds to compensate those harmed by their actions where they are moral agents. The desirability of AI holding such funds likely depends on personhood questions List addresses elsewhere. Absent personhood, corporate developers will most likely hold the funds. But that is straightforward corporate responsibility.

While List seeks "future-proofed" (1240) regulations, regulations for a far future should not come at the cost of tremendous benefits today and this regulation may also prove undesirable long-term. Consider a case from the explicitly 'high-stakes' medical sphere. Even the best healthcare systems that exist today are rife with iatrogenic injury (viz., provider-caused error) and misdiagnosis. If technical trends continue, medical AI will only perform better over time and yet may not be fully moral. It is non-obvious that such highly-beneficial AI should be prohibited ex-ante where administrative bodies can and will continue to check their performance and can take them off the market where they pose risks absent recognition thereof as 'moral' actors.[2]

List would only permit such AI where others face strict liability for its use but the transfer agreement requirement is itself problematic. We recognized AI and group agency to better track ideal allocation of responsibility. List worried that we could not properly allocate it by focusing on individual responsibility. It then seems odd to require non-ideal allocations by fiat. We

[2] Topol (2019) helpfully summarizes medical AI trends. I discuss regulatory checks in [redacted for review].

assume responsibility, but not accountability, where AI agents are not moral agents. We also assume that the corporate developer, its members, AI users, etc. are not fully responsible. This is why gaps arise. We do not, however, fill responsibility gaps for the mere sake of it (as List admits when discussing hurricane-caused damage (1227)). We fill gaps to ensure *proper* accountability for actions. But if no one is most fit to be held accountable, picking one out as having to bear more responsibility to fill a gap is problematic. Whatever one's general views on contentious strict liability doctrines, we should not adopt it just to make *someone* responsible.

Strict liability actually risks injustice in the case at issue where amoral AI is generally desirable such that we want 'backup' means of permitting their use. For instance, increased mortality and morbidity alone, let alone increased efficiencies, could justify AI that produces more accurate, efficient medical diagnosis. Society at large will benefit from them. We may want to avoid responsibility gaps for the use thereof. But it is problematic if the cost of such beneficial tools falls on those who, ex hypothesi, lack full moral responsibility in the relevant sense.

Strict liability, then, creates *too much* responsibility. People who are explicitly not fully morally responsible for actions are deemed fully legally responsible just so society can deem *someone* responsible. While private law should provide a means of ensuring that those harmed are compensated, this goal should not come at the expense of unfair distributions of burdens.

This problem is acute in List's high-stakes cases. The scope of the harms in less controversial strict liability cases is much narrower. I am unaware of harms at the scale of errant AI-enabled military strikes where we already plausibly hold someone strictly liable. Even large-scale food quality errors or speeding violations on a highway do not produce as much harm. Where we already depart from ideal responsibility allocations, then, we do so at a much smaller scale.

Potential disanalogies between the group and AI cases further undermine List's argument. There are, e.g., many cases of existing adaptive machine learning tools where no one has 'control' over an AI tool equivalent to board member control over a corporation. Adaptive tools will change over time. Programmers cannot, again, fully predict how it will change and loses contact with it that would permit supervision once it is on the market. A hospital purchaser may not, in turn, be able to catch every development. This requires a strong regulatory approach that can identify issues throughout the tool's lifecycle. It may not require that any one person be responsible for all harms. Indeed, no one seems like an appropriate choice for strict liability in such a case.

There are, moreover, no pre-theoretical/pre-institutional reasons to identify anyone most appropriate for acquiring full responsibility in AI cases parallel to group ones. There is a constitutive relationship between a group and its members. Some members are then more responsible for the group by institutional design. They are at least better candidates for full responsibility where/if strict liability is plausible. No such relationship exists in AI cases. In strict corporate liability cases, in turn, someone agreed to take on legal responsibility distinct from their pre-institutional moral responsibility. That has not yet happened in AI cases. We are now asked to find someone should take this role. No one seems distinctly appropriate.

One may still be tempted to state that requiring you to accept a stipulation that would deem you liable to gain permission to create AI and holding you liable in those cases is fair.[3] But this too requires identifying someone who could justifiably accept this role without intuitively problematic results. The most plausible construction of a strict liability regime for autonomous AI I can envision simply applies the corporate case: those who agreed to be responsible for a company are held responsible for its outputs. List bears the onus of establishing a plausible alternative. Without it, the reasons justifying transfers are group agency-based, not AI-specific. We are no longer discussing parallel arguments in that case but specifying group agency. Extant corporate laws then likely cloud our judgments.

Liability transfer requirements are also unnecessary to compensate relevant harms. A no-fault insurance scheme for medical AI could, e.g., ensure that those harmed by it received funds to address their harms (Mahila et al. 2021). Everyone will be compensated. No one will be liable for widespread harms for which they are not morally responsible. This produces List's desired outcome without perverse liability attributions. Parallels may not be available in all settings. Yet even one example establishes that List's condition should not apply to all high-stakes cases.

Conclusion

List's condition is undesirable. Moral engineering is neither necessary nor sufficient for appropriately balancing AI's potential benefits and risks of harm. Moreover, liability transfer agreements are both problematic and unnecessary to compensate victims of AI-enabled harms.

Bibliography

Himmelreich, Johannes. 2019. "Responsibility for Killer Robots." *Ethical Theory and Moral Practice* 22(3):731-747.

List, Christian. 2021. "Group Agency and Artificial Intelligence." *Philosophy & Technology* 34:1213-1242.

Mahila, George et al. 2021. "Artificial Intelligence and Liability in Medicine." *Milbank Quarterly* 99(3): 629-647.

McCradden, M. et al. 2020. "Ethical Limitations of Algorithmic Fairness Solutions in Health Care Machine Learning." *Lancet Digit Health* 2(5):e221-e223.

Topol E. 2019. "High-Performance Medicine." *Nature Medicine* 25:44-56.

Vyas, D.A. 2020. "Hidden in Plain Sight." *N Engl J Med* 383:874-882.

[3] [Redacted] made this point.