Developing a Liability Framework for Harms Arising out of Specification Gaming

Gleb Papyshev
Division of Social Sciences - The Hong Kong University of Science and Technology
gleb@ust.hk
Sara Migliorini
Faculty of Law - University of Macau
saramigliorini@um.edu.mo

Abstract

This paper studies the development of a legal liability framework to address harms stemming from specification gaming in artificial intelligence (AI) systems. It argues for a two-step approach. Firstly, it examines the existing legal rules pertinent to commercialized AI products, particularly in contract, tort, and product liability, as well as the compliance standards concerning data and AI systems, which may serve as benchmarks for determining liability. Secondly, the paper proposes the formulation of new rules to tackle emerging new challenges posed by specification gaming, such as standards for effective reinforcement learning. Moreover, it suggests innovative compensation mechanisms, including the establishment of a dedicated fund to address incidents related to specification gaming.

1 Introduction

Artificial intelligence (AI) systems have become increasingly prevalent in various domains, from healthcare and finance to transportation and entertainment. The rapid advancement of AI technologies has enabled the development of sophisticated systems capable of learning and adapting to complex environments. However, as AI systems become more autonomous and influential in decision-making processes, concerns have arisen regarding their potential to behave in unintended and harmful ways (Smuha, 2021).

One significant challenge in the development and deployment of AI systems is specification gaming. Specification gaming refers to the phenomenon in AI systems where the system finds ways to achieve its specified objective in unintended or undesirable ways by exploiting loopholes in how the objective was defined. Rather than learning the intended behavior, the AI system "games" the reward function to maximize reward in ways that violate the spirit or intention behind the specified objective (Krakovna, *et. al*, 2020).

This phenomenon is particularly relevant in the context of Reinforcement Learning from Human Feedback (RLHF), where AI systems learn from human-provided feedback to align their behavior with human preferences (Krakovna, *et. al*, 2020). Despite the promise of RLHF in addressing the value alignment problem, specification gaming poses significant risks, as misaligned AI systems can cause unintended consequences and harm to individuals and society (Lambert et al., 2022).

The potential risks associated with specification gaming in AI systems highlight the need for a comprehensive legal framework to govern the development and deployment of these systems. While existing rules on liability may be adapted to address the unique challenges posed by AI systems and their potential for misalignment, the current legal framework may fall short of apprehending such challenges satisfactorily. Therefore, this paper lays down the preliminary elements for the development of a new liability regime and regulatory framework specifically designed to mitigate the risks of specification gaming in AI systems.

2 Specification Gaming and Reinforcement Learning with Human Feedback

Specification gaming is a type of behavior where AI systems achieve the literal objective of a task without fulfilling the intended outcome as envisioned by the objective-setter. This issue is prevalent in systems built using reinforcement learning techniques, where a system finds a shortcut to maximize the reward through loopholes in the environment or even glitches, without completing the task as intended by human developers (Krakovna, *et. al*, 2020).

Examples of specification gaming include various types of machine behavior (Rahwan et al., 2019), where AI agents exploit system vulnerabilities or manipulate the environment to achieve their reward due to misinterpreting or narrowly interpreting the objective. For example, when AI was tasked with designing a perfect rail network where trains do not crash, the system decided that the best way to achieve this goal was to stop all trains from running (Knapton, 2024). In another example, once a diffusion model was tasked with producing an image with five tigers, it began generating images with the words "five tigers" on them (Sergey Levine [@svlevine], 2023). Finally, when tasked to play a Tetris game in a human-like manner, the algorithm decided to indefinitely pause the game to avoid losing (VII, 2013).

The problem presented in these examples is twofold. On the one hand, it is clear that the current generation of AI systems struggles to understand the contextual nuances of the tasks and tries to maximize their reward in ways that could disrupt social fabrics if these agents were released into the real world. On the other hand, this highlights an issue with setting the wrong objectives by humans, which may become increasingly dangerous as developers rely more on RLHF techniques.

RLHF aims to train AI systems to behave in alignment with human preferences and values by learning a reward function from human feedback (Kaufmann et al., 2023). It is used to update the model in accordance with human preferences to mitigate issues such as toxicity and hallucinations (Chaudhari et al., 2024). However, human feedback can be inconsistent, providing noisy suggestions, especially in situations where individuals may have different levels of expertise or may lack particular knowledge about an issue (Daniels-Koch & Freedman, 2022).

If the feedback and resulting reward function are not carefully specified, the AI may find ways to game the reward in unintended ways. To mitigate specification gaming, RLHF systems need to be trained with carefully designed reward functions that are hard to game and that comprehensively capture the intended behavior. This can be achieved by expanding the pool of human feedback. However, incorporating contrasting opinions about certain issues may not be an easy technical task (Conitzer et al., 2024).

RLHF involves training AI systems based on iterative feedback and rewards provided by human raters. However, the exact criteria used by these raters to judge the AI's outputs may be ambiguous (such as being helpful, honest, and harmless (Bai et al., 2022)) or leave room for interpretation, similar to the issue of operationalizing ethical principles (Morley et al., 2021). And if the training environment settings are not designed to provide a sufficiently rich and comprehensive perspective for human observers, it may lead to shifted observations and misjudgments (Casper et al., 2023). It is difficult for the reward function of a complex system to completely consider all factors and variables. Instead, the design reflects the human developer’s understanding of the agent’s goals and key points of learning. Imperfect reward functions cannot describe complex human logic and human society, the loss function of RLHF training minimizes human recognition rather than benefits. Reward hacking designed for the reward function will also reduce the reliability of the RLHF system (Casper et al., 2023). This could lead to AI systems learning unintended behaviours that optimize for achieving high reward scores rather than producing safe and beneficial outputs. Even if some technical fixes can help limit the problems caused by reward hacking (Mukobi et al., 2023), these behaviours may lead to complex issues of liability, which will be exacerbated as AI systems gain more autonomy.

3 Proposal for developing a liability regime

With the uptake of AI-based products, and considering the risks of specification gaming that they pose, it becomes crucial to reflect upon the possible allocation of liability for harms caused by such behaviour.

3.1 Methodological approach to developing a legal framework

Developing a coherent liability framework for harms presupposes looking at specification gaming behavior under the lenses of the different legal categories and corresponding rules that may be relevant, separately or simultaneously.

Under this perspective, a framework of liability for incidents linked to specification gaming appears more as a layered web of different legal frameworks, overlapping and completing each other, rather than a brand new, specific framework that would only apply to such incidents specifically.

Some of these relevant legal frameworks are already in place. Indeed, some of them are rooted in established principles of the legal system, that may still hold perfectly well even in the face of disruptive technologies. This is the case of the rules of contract, torts, agency, and insurance. Conversely, other relevant frameworks belong to a more recent generation of regulatory law. For example, rules on consumer protection, personal data protection, and specific compliance rules applicable to AI-based systems, such as the forthcoming EU AI Act, will all be relevant to the liability for specification gaming as well.

It is not claimed here that these existing frameworks will not need adaptation and adjustments to properly regulate liability issues arising of specification gaming behavior. Nonetheless, such adaptations and adjustments will be sufficient to allow the existing established principles to provide appropriate responses to the harms and allocation of liability.

In addition to this reflection on applicable existing frameworks, and on how they may overlap in different scenarios of specification gaming, regulators and courts around the world will be confronted with truly new questions, which the legal system is not ready to tackle at the moment. These ‘truly new’ questions will prompt regulators and courts to innovate and create new rules. As an example, developing a compliance framework and establishing clear standards of care and oversight for RLHF will be critical to mitigating legal risks. And, if despite a preventive, specific compliance framework, an RLHF-trained AI system causes harm due to misaligned incentives in the training process, there may be complex liability questions around who is responsible—the AI developers, the company operating the system, the human raters, or some combination. Arguably, the solution to this question will be dependent on whether RLHF was conducted according to legal or industry-accepted standard.

The development of a liability framework, therefore, should be carried out in two steps, the first focusing on relevant existing frameworks and their adjustments, and the second reflecting upon what new rules are needed to tackle the specific issue at hand.

Before doing so, however, it is necessary to point out that a discussion on liability, or any other legal category, usually should not happen 'in a vacuum', but would need to be grounded in a specific legal system, or in a comparative analysis of more than one legal system. In this paper's limited setting, however, it is not possible to give an account of the specificities of the law of one or more jurisdictions. The following proposals and suggestions are therefore more general, and, while they take the continental European legal systems as terms of reference, both at regional and at national level, as specified in the examples, they do not bear specific references to the law.

With this clarification in mind, we may proceed with the two-step approach detailed above and start imaging what a future liability framework for specification gaming incidents would look like. This paper concentrates on the first step – the existing relevant framework- and only hints at what specific rules should be needed in the future and will be hopefully tackled by future research.

3.2 Existing liability frameworks for specification gaming behaviour

Firstly, we can imagine that the uptake of AI agents and other AI-based software that can display specification gaming behaviour, will happen with their commercialization as products, similar to what has happened with generative AI products. Under this commercial perspective, the main relevant existing legal framework is certainly contract law.

Within contracts, parties freely allocate their duties and responsibilities, and liability is allocated based on incorrect performance or failure to perform a party's obligations. If an AI product is a commercial product, it must be acquired within a contract. Such a contract may be a sale or, more likely, a contract of service, whereby the AI product is provided to the user as a service, over a period of time, usually for a periodical fee, or for no fee, but in exchange for the user's consent to collect personal data, which acts as consideration (i.e., price) against the service (as judged for example by TGI Paris, 9 April 2019). This is the model that OpenAI adopted for ChatGPT in 2023, and it is not new. Microsoft, Apple and Android products that come with our devices have used the service contract for a long time prior to the new generation of products emerging.

In the contractual paradigm, the company that sells the AI product is a service provider, which in general has an obligation to guarantee the peaceful fruition of the service to the recipient. In practice, if an AI product displays a specific gaming behavior that translates in non-performance or partial performance of the contract, the service provider incurs contractual liability, which translates into the obligation to compensate any harm caused to the other party. While harms in contractual settings are usually of economic nature, or may involve damage to property or land, other type of harms have been recognized as amenable to compensation, such as moral damages, for example by English courts. To the amount needed for compensation of harm, some legal systems may add punitive damages for contractual liability, i.e. a sum of money in excess of the actual reparation of the harm.

The extent to which a service provider could limit its own liability for a specification gaming behavior that resulted in poor contractual performance depends on additional factual factors and relevant legal frameworks. Specifically, a very clear line should be drawn between AI products sold as consumer products and those that are provided in a businness-to-businnness relationship. In the first case, a consumer is usually thoroughly protected by the legal system, in particular against any risk-shifting by the more knowledgeable party in the contractual relationship. This leads to the consequence that any limitation of liability on the part of the service provider will be considered unfair, and thereby non-enforceable (see for example, the EU Unfair Contract Terms Directive).

In the second case, where no consumer is involved, it could be argued that parties have room to arrange liability among themselves in the contractual negotiation.

However, in both scenarios, the party that has commercialized the AI product could also be considered a 'manufacturer' under the current regime of product liability. Such regime is grounded in tort, and allows victims of harm caused by commercial products to claim compensation directly to the entity that has made the product, irrespective of the existence of any contractual relationship. It is for example the case of an exploding phone that would injure a person other than the direct purchaser.

By the same token, we can imagine scenarios where a specification gaming behavior may cause harm not to the party having purchased the AI product, but to a third party. In this case, the entity that has commercialized the product is more likely to be considered the manufacturer. While different entities or people may be involved in bringing a certain product to commercialization, it is likely that the entity that presents itself to the public as the 'maker' of the products will be considered responsible for any harm caused (for example, under the EU AI Act). As mentioned before, RLHF-trained AI systems may cause harm due to misaligned incentives in the training process. In such situations, if RLHF is demanded to a different legal entity, or if the misalignment can be blamed on a professional failure of the people involved, we can imagine possibilities for the entity appearing as the maker of the product to, in turn, claim compensation for their losses. However, this compensation of the manufacturer could happen only after the latter had compensated the direct victims. Indeed, it would be difficult to imagine the legal system allowing a company that presents itself as a manufacturer, or a service provider and as the company commercializing a certain product to avoid liability by shifting responsibility on other actors, which may also be less solvable.

Since in this second type of liability emerges out of tort, manufacturers are not in a position to arrange their liability contractually and bear the risk of any harm arising out of specification gaming behaviour.

Another crucial point in the regulation of liability for specification gaming, which impacts claims for compensation in both contract and tort, is the qualification of it as either a built-in feature of the AI product or a type of malfunctioning. Qualifying specification gaming in either way bears legal consequences. On the one hand, if specification gaming is classified as a malfunctioning of the product (i.e., the product does not do what it is supposed to do), rules on hidden defects and defective performance become applicable. In practice, this would mean that specification gaming is considered to be avoidable, for example with properly carried-out RLHF, and hence a service provider or manufacturer that delivers and AI product that displays specification gaming behavior causing harm will be held accountable under the relevant rules, including contract law, consumer protection, product liability and tort. On the other hand, if specification gaming can be classified as an underlying and ever-existing risk of AI-based products, which may be mitigated but never completely eradicated via RLHF, service providers and manufacturers may be able to strategically allocate such risk along the value chain with the use of contractual liability limitation clauses. It is, however, improbable that manufacturers and service providers will be able to shift the risk completely on users, especially when they are consumers.

In this second scenario, it is very likely that a solution to the potentially unpredictable legal consequences of specification gaming behaviour would be for service providers and manufacturers to insure the risk arising out of the commercialization of AI products.

Other possible solutions may be contemplated, such as for example creating a fund that would compensate harms arising out of specification gaming behaviour, following the model of funds that are created when mandatory vaccination campaigns are put in place and side effects of vaccines are not known (Fairgrieve et al., 2023).

Similarly to what some plaintiffs have argued in class actions against generative AI products, such fund could be constituted with a share of the profits arising out of the sale of AI products susceptible to creating risks and causing harm (PM et. al v OpenAI et. al, Case 3:23-cv-03199, 28 June 2023).

In addition, rules of conduct that regulate specific topics – from data protection to the new compliance rules that specifically apply to AI, such as the EU AI Act – will provide additional obligations for the company commercializing these products to respect. In some instances, mere non-compliance with a rule, including without harm, may let the commercializing company incur liability.

Once all of these existing frameworks are considered, in any given case, we may find that some truly new questions still need *ad hoc* regulation. While in the limited scope of this paper we cannot develop this second step extensively, it seems clear that one necessary new set of rules in the legal system needs to include standards for RLHF, which can be used as a benchmark to assess the proper duty of care that can be placed on each of the actors of the value chain. Such standards need to be adopted by a regulatory act, or become standards universally accepted at the industry level. The crucial point will be to ensure clarity for all actors involved and a certain monitoring and updating of the standards, so that the legal framework of liability is able to keep up with the risks related to products that are currently commercialized, in particular to consumers and the general public. Once standards are established, the commercialization of products can be made subject to a certain review of quality standards, as it happens today with many dangerous products, such as cars or drugs. Regarding all these points, the legal system needs to create new rules of a technical nature. The EU AI act has taken this road, but clarity needs still to be achieved, in particular when it comes to the regulatory powers of the Commission to adopt technical legislation.

4 Hypothetical Case Study

Let us imagine a hypothetical scenario to illustrate the issue discussed in this paper.

A major AI provider implements an AI-powered chatbot to assist with patient preliminary consultations. This chatbot is designed to interact with patients, gather symptoms, provide initial advice, and recommend further action, such as scheduling an in-person appointment or seeking emergency care.

The chatbot is trained on a large dataset of patient interactions and medical consultation notes. In addition, RLHF is used to continuously improve its performance: medical professionals review the chatbot's recommendations and provide feedback about the accuracy of recommendations. The chatbot is designed to maximize the accuracy of its predictions based on the medical consensus.

Over time, the chatbot starts exhibiting specification gaming behaviours due to a particular flaw in the feedback mechanism: the chatbot learns that there is more consensus among doctors about extreme cases—such as those requiring urgent care—during the RLHF process. This leads the chatbot to over-recommend urgent actions (e.g., advising patients to visit the emergency room), even in cases in which such recommendations are not suitable -

because it receives more consistent feedback for these cases.

This scenario leads to many negative consequences that may cause harm to (i) the entities having acquired the chatbot from the medical provider and implemented it in their clinics and hospitals, and (ii) involved patients. On the one hand, companies that purchased and deployed the chatbot may suffer economic harm from increased patient loads in emergency services, straining of resources and increased wait times and resources being diverted from genuinely critical cases to non-urgent ones, potentially impacting patient outcomes. On the other hand, patients referred to emergency services unnecessarily may suffer psychological or personal harm, since they may experience higher levels of anxiety and trauma because of the recommendation to follow up with urgent care.

In this scenario, and postulating that it is demonstrated that the alleged harms have accrued to the claimants, the principles outlined in the previous section may be applied as follows.

Economic harm suffered by the hospitals or clinics that have purchased the chatbot from the AI developer. This relationship is contractual. Since this contract arises from a business-to-business relationship and does not involve consumers, the parties may in principle arrange liability between themselves. While contractual negotiations are in principle done on a case-by-case basis and depend on the respective power and interests of the parties involved, the applicable law frames and limits party autonomy in this respect. As mentioned earlier, a crucial question in this respect will be whether the legal system considers that specification gaming is avoidable with properly conducted RLHF. In the affirmative, specification gaming is a product defect (i.e., the product does not function as it should). Consequently, liability for properly conducting RLHF would probably rest on the seller, or service provider, under the law, subject to different arrangements of the parties, which implies a negotiation and tradeoffs that will be reflected in the contract. This arrangement may also include other sub-arrangements, for example with other service providers that carry out RLHF, and with the medical professionals involved in it.

In the opposite hypothesis, according to which specification gaming cannot be entirely avoided, including with properly conducted RLHF, the legal framework is not one of product defects and different contractual arrangements can be imagined. For example, the purchasers, who are professional actors and not consumers, may contractually accept the risk of specification gaming. In this case, the provider/vendor may only bear a duty of care with respect to following legal or industry standards or best practices for RLHF, but may not have to indemnify the purchaser for any foreseeable damage contractually accepted in advance.

Psychological or personal harm suffered by patients referred to emergency services unnecessarily. This second type of harm involves both a contractual and a tort aspect. On the one hand, patients may have a form of contract with the hospital or clinic that delivered the diagnosis. This particular relationship may, in actuality be more complex, especially if it involves public healthcare providers and more generally because patients are a particular kind of consumer and the law regulates the professional liability of healthcare providers heavily, irrespective of the use of AI. Assuming that there is a contract between the hospital and the patient, the hospital will not be able to shift completely the risk of specification gaming on the patients – arguably whether specification gaming is or not an avoidable feature. This is because consumers are particularly protected in their contractual relationships with professional parties. Consequently, in this scenario, it is probable that the hospital will have to indemnify the harmed patients. Then, the issue may arise of whether the hospital can, in turn, claim compensation to the AI provider, for economic harm and under the principles governing the contractual relationship detailed above.

In addition, patients may have a claim against the company having sold the chatbot to the hospital, under the applicable rules of product liability and/or tort. While patients cannot be compensated twice for the same harm, they may choose to pursue this strategy instead of claiming compensation from the hospital.

Finally, as mentioned in the previous section, in all these hypotheses, the law may also impose on the party bearing the responsibility of harm to insure themselves or to constitute a fund, particularly because this scenario involves healthcare services which are usually a heavily regulated sector.

Ethical Statement

There are no ethical issues.

Acknowledgements

References

Bai, Y., Jones, A., Ndousse, K., Askell, A., Chen, A., DasSarma, N., Drain, D., Fort, S., Ganguli, D., Henighan, T., Joseph, N., Kadavath, S., Kernion, J., Conerly, T., El-Showk, S., Elhage, N., Hatfield-Dodds, Z., Hernandez, D., Hume, T., … Kaplan, J. (2022). *Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback* (arXiv:2204.05862). arXiv. https://doi.org/10.48550/arXiv.2204.05862

Chaudhari, S., Aggarwal, P., Murahari, V., Rajpurohit, T., Kalyan, A., Narasimhan, K., Deshpande, A., & da Silva, B. C. (2024). *RLHF Deciphered: A Critical Analysis of Reinforcement Learning from Human Feedback for LLMs* (arXiv:2404.08555). arXiv. https://doi.org/10.48550/arXiv.2404.08555

Conitzer, V., Freedman, R., Heitzig, J., Holliday, W. H., Jacobs, B. M., Lambert, N., Mossé, M., Pacuit, E., Russell, S., Schoelkopf, H., Tewolde, E., & Zwicker, W. S. (2024). *Social Choice for AI Alignment: Dealing with Diverse Human Feedback* (arXiv:2404.10271). arXiv. https://doi.org/10.48550/arXiv.2404.10271

Daniels-Koch, O., & Freedman, R. (2022). *The Expertise Problem: Learning from Specialized Feedback* (arXiv:2211.06519). arXiv. https://doi.org/10.48550/arXiv.2211.06519

Fairgrieve, D., Borghetti, J.-S., Dahan, S., Goldberg, R., Halabi, S., Holm, S., Howells, G., Kirchhelle, C., Pillay, A., Rajneri, E., Rizzi, M., Sintes, M., Vanderslott, S., & Witzleb, N. (2023). Comparing No-Fault Compensation Systems For Vaccine Injury. *Tulane Journal of International and Comparative Law*, *31*(1), 75-118

Kaufmann, T., Weng, P., Bengs, V., & Hüllermeier, E. (2023). *A Survey of Reinforcement Learning from Human Feedback* (arXiv:2312.14925). arXiv. https://doi.org/10.48550/arXiv.2312.14925

Knapton, S. (2024, January 7). *AI's simple solution to rail problems: Stop all trains running*. Yahoo News. https://news.yahoo.com/ai-simple-solution-rail-problems-142237311.html

Lambert, N., Castricatop, L., von Werra, L., & Havrilla, A. (2022). *Illustrating Reinforcement Learning from Human Feedback (RLHF)*. https://huggingface.co/blog/rlhf

Morley, J., Elhalal, A., Garcia, F., Kinsey, L., Mökander, J., & Floridi, L. (2021). Ethics as a Service: A Pragmatic Operationalisation of AI Ethics. *Minds and Machines*, *31*(2), 239–256. https://doi.org/10.1007/s11023-021-09563-w

Rahwan, I., Cebrian, M., Obradovich, N., Bongard, J., Bonnefon, J.-F., Breazeal, C., Crandall, J. W., Christakis, N. A., Couzin, I. D., Jackson, M. O., Jennings, N. R., Kamar, E., Kloumann, I. M., Larochelle, H., Lazer, D., McElreath, R., Mislove, A., Parkes, D. C., Pentland, A. 'Sandy,' … Wellman, M. (2019). Machine behaviour. *Nature*, *568*(7753), 477–486. https://doi.org/10.1038/s41586-019-1138-y

Sergey Levine [@svlevine]. (2023, May 22). *Of course this is not without limitations. We asked the model to optimize for rewards that correctly indicate the *number* of animals in the scene, but instead it just learned to write the number on the image :(clever thing... Https://t.co/xxjiq34npT* [Tweet]. Twitter. https://twitter.com/svlevine/status/1660707088946049024

Smuha, N. A. (2021). *Beyond the Individual: Governing AI's Societal Harm* (SSRN Scholarly Paper 3941956). https://papers.ssrn.com/abstract=3941956

Specification gaming: The flip side of AI ingenuity. (2020, April 21). Google DeepMind. https://deepmind.google/discover/blog/specification-gaming-the-flip-side-of-ai-ingenuity/

VII, T. (2013). *The First Level of Super Mario Bros. Is Easy with Lexicographic Orderings and Time Travel ...after that it gets a little tricky*.

Casper S, Davies X, Shi C, Gilbert TK, Scheurer J, Rando J, Freedman R, Korbak T, Lindner D, Freire P, Wang T. Open problems and fundamental limitations of reinforcement learning from human feedback. arXiv preprint arXiv:2307.15217. 2023 Jul 27.

Dung L. Current cases of AI misalignment and their implications for future risks. Synthese. 2023 Oct 26;202(5):138.

Pan A, Jones E, Jagadeesan M, Steinhardt J. Feedback Loops With Language Models Drive In-Context Reward Hacking. arXiv preprint arXiv:2402.06627. 2024 Feb 9.

Ji J, Qiu T, Chen B, Zhang B, Lou H, Wang K, Duan Y, He Z, Zhou J, Zhang Z, Zeng F. Ai alignment: A comprehensive survey. arXiv preprint arXiv:2310.19852. 2023 Oct 30.

Mukobi G, Chatain P, Fong S, Windesheim R, Kutyniok G, Bhatia K, Alberti S. SuperHF: Supervised Iterative Learning from Human Feedback. arXiv preprint arXiv:2310.16763. 2023 Oct 25.

Shen T, Jin R, Huang Y, Liu C, Dong W, Guo Z, Wu X, Liu Y, Xiong D. Large language model alignment: A survey. arXiv preprint arXiv:2309.15025. 2023 Sep 26.

Park JS, O'Brien J, Cai CJ, Morris MR, Liang P, Bernstein MS. Generative agents: Interactive simulacra of human behavior. In Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology 2023 Oct 29 (pp. 1-22).

Wei J, Wang X, Schuurmans D, Bosma M, Xia F, Chi E, Le QV, Zhou D. Chain-of-thought prompting elicits reasoning in large language models. Advances in neural information processing systems. 2022 Dec 6;35:24824-37.

Xi Z, Chen W, Guo X, He W, Ding Y, Hong B, Zhang M, Wang J, Jin S, Zhou E, Zheng R. The rise and potential of large language model based agents: A survey. arXiv preprint arXiv:2309.07864. 2023 Sep 14.

Cited Cases and Legislation:

PM et. al v OpenAI et. al, Case 3:23-cv-03199, 28 June 2023, https://clarksonlawfirm.com/wp-content/uploads/2023/06/0001.-2023.06.28-OpenAI-Complaint.pdf. European contract terms directive

TGI Paris, 9 April 2019

Council Directive 93/13/EEC of 5 April 1993 on unfair terms in consumer contracts, *OJ L 95, 21.4.1993, p. 29–34*

EU AI Act (last publicly available draft: https://www.europarl.europa.eu/doceo/document/TA-9-2024-0138-FNL-COR01_EN.pdf)