

Generative AI and the Rule of Law^{*}

Pompeu Casanovas^{1,*}, Mustafa Hashmi² and Marta Poblet³

¹IIIA-CSIC and UAB Institute of Law and Technology (IDT-UAB)

²UAB Institute of Law and Technology (IDT-UAB) and La Trobe University

³RMIT University, Melbourne

Abstract

Artificial Intelligence impacts the way we think and behave. It is already changing both, and it is commonplace by now that Large Language Models (LLMs) will be widely adopted. Existing platforms, including Google, are embedding AI into their search engines to optimise their results. This creates a new set of scenarios where the Turing test is being challenged. However, far from answering the questions brought by practical philosophy, these developments raise new questions as they foster the interest for computational ethics, the modelling of values and principles, and the emergence of legal ecosystems, i.e., the establishment of complex hybrid or symbiotic environments able to turn norms into living rules on real time. In this exploratory paper, we describe the emergence of LLMs and Multimodal Foundation Models (MFMs) and make some suggestions about how they can be used for regulatory purposes and modelling the rule of law.

Keywords

Large Linguistic Models (LLMs), Multimodal Foundation Models (MFMs), ChatGPT 3.5, ChatGPT-4, Claude, Regulatory models, Rule of law

1. Introduction


Chat Generative Pre-Trained Transformers 3 and 4 (ChatGPT-3.5, GPT-4) are AI tools developed by OpenAI, the AI company focusing on producing tools for the automated generation of images and natural language understanding (NLU). NLU includes multiple tasks, such as textual entailment, semantic similarity assessment, document classification, natural language inference, question answering, and common-sense reasoning. As part of their solutions, Open AI figured out a “generative pre-training of a language model on a diverse corpus of unlabelled text, followed by discriminative fine-tuning on each specific task” [1]. After the creation of some prototypes, OpenAI finally came up with the last LLM of this series, which immediately gained worldwide attention due to the astonishing quality of the answers. Not only syntactically consistent with the grammar of the proposed natural language, these answers are intended to be semantically coherent and pragmatically cohesive. Readers can infer a general meaning


Proceedings of Artificial Intelligence Governance Ethics and Law (AIGEL), Reviewed, Selected Papers. November 02 - December 19, 2022, Barcelona, Spain

*Corresponding author.

✉ Pompeu.Casanovas@uab.cat (P. Casanovas); M.Hashmi@latrobe.edu.au (M. Hashmi); marta.pobletbalcell@rmit.edu.au (M. Poblet)

ORCID 00000-0002-0980-2371 (P. Casanovas); 0000-0002-6376-082X (M. Hashmi); 00000-0002-0026-989X (M. Poblet)

 © 2022 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

from the generated text as they experience the sense of fluent talk that usually comes with meaningful conversational exchanges.

How is this possible? While some of the architectural innovations are being shared publicly, “there is a paucity of information about the development, deployment and commercialisation of these models and the applications and services based upon them” [2]. Other considerations point out that it could well be that GPT-4 does not hold a single model but an $8 \times 220\text{B}$ mixture of experts (MoE) model, i.e., 8 smaller models tied together and using a technique first implemented by Google in 2021 [3]. All this still needs confirmation. One of the last revelations were disclosed to the New York Times by members of Anthropic, a venture company lab launched in 2021 by a group of former employees of OpenAI “who grew concerned that the company had gotten too commercial” [4]. Anthropic then released another Chatbot, Claude, following what they call *Constitutional AI* principles, i.e. embedding automated constitutive values to make the system compliant with ethics and the law [5]. We will review this assertion below.

LLMs and Multiple Foundations Models (MFMs) recognise patterns in data and produce sophisticated answers based on those patterns [2]. A foundation model “is a type of machine learning model that learns from a wide range of data using self-supervision at scale” [6]. Transformers’ innovation parallels the one on language processing, but transformers analyse all tokens in a given body of text simultaneously, rather than in sequence [ibid, 6]. Transformers rely on an AI mechanism known as *attention*, enabling “a model to consider the relationships between words and to determine which words and phrases in a passage are most important to pay attention to” [ibid, 6].

MFMs and LLMs lean on deep learning but, unlike other LLMs, ChatGPT targets the ability to learn effectively from raw text without any need of manual labelled data, i.e., without the need of prior annotation processes. Clearly, the main goal seems to be to reach a universal representation applicable to a wide range of tasks—e.g., machine translation, document generation, or syntactic parsing—without any prior domain specification. The founders use the expression “general task-agnostic model” to refer to this feature of unsupervised pre-training as a special case of semi-supervised model [1]. The training procedure has two steps, in which supervision is reserved for fine-tuning on a target task. The first stage is learning a high-capacity language model on a large corpus of text. The second one adapts the model to a discriminative task with labelled data, i.e., “transferring a variable number of layers from unsupervised pre-training to the supervised target task” [ibid, 1].

Many users in a large variety of fields have already shared their feelings and some test results of ChatGPT and GPT-4 performances. They are generally positive, but many have pointed out its limitations while expressing some ethical concerns about its use. For example, in education, it has been argued that “the aid of ChatGPT did not necessarily improve the quality of students’ essays” [7]. Perkins [8] shows a similar view. In medicine, after a serial of five prompted questions on surgical management of knee osteoarthritis (OA), Seth et al. [9] conclude that the answer was relevant, albeit surface-level and with a set of essential mistakes showing that results were not totally accurate and the system “lacked the analytical ability to dissect for important limitations about knee OA, which would not be conducive to potentiating creative ideas and solutions in orthopaedic surgery” [ibid, 9]. In corporate analysis, Neumann and Wolczynski [10] has shown “that benefits from AI-assisted fact-checking may be unequally received amongst online communities, implying that mis- and disinformation that disproportionately impacts

minority communities may indeed be falling through the cracks.” The presence of biases and side effects has been generally noticed as well.

As of July 2023, a cascade of experiments and testings have been already performed, GPT-3.5, GPT-4 and LLMs in general have been widely adopted in everyday life. As it was the case with search engines twenty years ago—transformers have come to stay. To be sure, teenagers have been using them to do their homework before widespread adoption at the beginning of this year. The same is happening with routine tasks in companies and organisations. If that is the case, LLMs can also help to model routine legal tasks (such as writing lawsuits) but, interestingly, LLMs can also assist in modelling more abstract legal foundations, such as the implementation of ethical principles, the enhancing of rights, and the formulation of the rule of law.

2. Plausible results from Plato’s cave

Not all uses are acceptable, and some hilarious mismatches have already occurred at the users’ level.¹ As it is well-known by now, false answers or “hallucinations” are a common feature ChatGPT. Based on probabilities, its research space deals with signs and symbols, which is a semiotic dimension (Kubacka [11], as quoted by Adamson [12]), but not with truth. Names, events and processes should not be considered as reflecting any reality. It provides eerily human-like responses to user questions (or *prompts*). They cannot be deemed human responses, or human-linguistic responses. Hence, from a linguistic perspective, they deal with meanings without reference or, better, with text in a mathematical space. This comes back to the difference between the semantic dimension of language (*meaning*) and its interpretation, use, and management in a specific context (*sense*). Interestingly, Kubacka uses the word ‘plausibility’. ChatGPT would make people experience “a plausibly sounding, non-existing parallel universe of phenomena”. Results sound not only probable, but *plausible*. What does it mean?

Plausibility is a complex notion because it means that something could be not just probable but possible. Its ambiguity lies on the double address of its driving: (i) to the world (making something true); (ii) and to the audience (producing a true effect or impact). The Merriam-Webster Dictionary states that *senses*, not *reason*, is the root of its etymological origins. It reads: “In fact, plausible comes to us from the Latin adjective *plausibilis*, meaning ‘worthy of applause’, which in turn derives from the verb *plaudere*, meaning ‘to applaud or clap’.”² So, this brings into play rhetoric and reasoning at the same time, in a kind of *metaxy* (*μεταξύ*), to use Plato’s word in the *Philebus* and the *Symposium*, expressing the existential tension between beliefs and truth, in-betweenness, in-Between.³ We can pull from one end of this epistemic rope or

¹A prototypical example of unmonitored or inaccurate usage of ChatGPT to write lawsuits has recently occurred. Cf. <https://www.theverge.com/2023/5/27/23739913/chatgpt-ai-lawsuit-avianca-airlines-chatbot-research>

²<https://www.merriam-webster.com/dictionary/plausible>

³The German philosopher and political scientist Voegelin [13, pp.251 and ff.] drew the attention to this notion of *metaxy*. According to Torres and Monserrat [14], eds. *Voegelin’s Political Readings: From the Ancient Greeks to Modern Times*. Routledge, 2021. “[...] as Voegelin formulated it in his unfinished *In Search of Order*, between the “things-reality” and the “it-reality” [Voegelin 13, CW, 18, 28-31]. Thus, “the experience,” states Voegelin, “is neither in the subject nor in the world of objects, but in-Between, and that means In-Between the poles of man and of the reality that he experiences” [Voegelin 13, CW, 34:98]. Man exists under an irreducible tension between the divine *νοῦς* that pulls us on the one side; and the human *νοῦς* (or reason) that moves us toward its origins thanks to the strength of the $\Psi\nu\chi\acute{\eta}$ on the other. As Voegelin would put it, there is no *ζήτησις* (search) without *έλκειν* – the pull

from the other one but very likely we will end up pulling it from both and at the same time. We should be able to make sense of the outer from this new inner Platonian cave to figure out how to proceed.

This is more than an analogy. Let's compare this perspective with some results recently delivered by the Microsoft team testing LLMs and GPT-4. They claim that GPT-4 is "is strikingly close to human-level performance", and "it could reasonably be viewed as an early (yet still incomplete) version of an artificial general intelligence (AGI) system". One of the experiments consists of generating Plato's Dialogues. The Microsoft team produced some hypothetical dialogues by Plato criticizing autoregressive language models written by ChatGPT and GPT-4 evaluation and they compared them.⁴ This is an impressive result, Plato-like, provided that we do not understand this exercise as a real reproduction of the ancient Greek philosophical style of writing and conveying meaning, i.e. as a philosophical piece of work. Users can pull meaning out of it, but this is unrelated with any historical evaluation of what a Plato's Dialogue consists of. Greek philosophy was related not to writing but to speaking and the paidetic intention of the speakers. It was linked to orality. A way of life.⁵ The exercise of reproducing dialectics as iterative autoregressive dialogues (texts) is just a contemporary recreation, an image, a simulacrum according to the old Scholastics. It is what it is, nothing else, and it cannot be criticised (nor assumed) from a historical, referential, or even cognitive point of view. Rather, it can be conceived as a useful engineering tool.

3. A rhetorical turn

The operation of making sense of symbolic expressions in a social environment, be they formulated in a formal, semi-formal or a natural language, raises the old theoretical problem of the semantic-pragmatic link. Commenting on the epistemic approach encouraged by DARPA, model induction, Adamson [12] points out that it is based on post-hoc reasoning: "Such inductive reasoning is consistent with the scientific method [12]. However, that method has a history of controls that are applied to build confidence in an uncertain, inductive, outcome".

Scientists are using LLMs and GPT4 for several routine tasks involving text generation, such as abstracts, memories, projects, summaries etc. But there are other tasks that the system can perform as well with acceptable results: "Researchers can use ChatGPT to debug and annotate code, translate software from one programming language to another and perform rote, boilerplate operations, such as plotting data" [17]. The same author (ibid.) offers six tips for better coding considering a few prudential rules: (i) choose small, discrete programming tasks (such as loading data); (ii) verify (to find bugs in coding); (iii) think safely (check for security vulnerabilities); (iv) iterate the queries (refining the answers); (v) anthropomorphise (treat chatbots as they were people, assuming roles, breaking the conversation into smaller pieces, etc.); (vi) embrace change.

These strategies and related tactics are not science but rhetoric in the classical sense, figuring out an audience, an interlocutor, choosing communication means as in a theatre play, and

understood as the experience of the power of attraction exercised by transcendence [13, CW, 12:281].

⁴See the full outcomes in Bubeck et al. [15, 10, Fig. 6.1].

⁵See the works by Hadot [16], among many others.

entering dialogue with her and with the user’s purpose and intention. For example, the inference of an intentional meaning from the results leans on the user’s side, it depends on the strategies and the actions taken by the user. Thus, as we contend below, prompting is of utmost importance, and this set of moves can be also performed by an artificial agent.⁶

Bell et al. [2] have identified the lifecycle stages of strategies throughout the generative AI model development from the user’s point of view: (i) Model pre-training (trained on large datasets to perform general and abstract tasks such as predicting subsequent words, in the case of LLM); (ii) Model fine-tuning for a specific application, requiring human intervention; (iii) Implementing input and output filtering (shaping the user inputs to a model); (iv) Pre-release testing (throughout the development process, and especially before release); (v) Release and distribution (choosing a range of strategies for making the system available); (vi) Post-release monitoring (developers will continue to monitor the way users interact with an application).

Hence, seeking intentional meaning is relevant; but there is a hidden interaction between users and developers not through human language but through a human-like language generation. It turns out that formulating prompts and refining them helps to assess intentional meaning. The whole process reminds what pragmatics, cognitive science and ethnomethodological disruptive experiments (‘breaching rules’) revealed more than sixty years ago: Humans project and ‘find’ linguistic meaning by figuring out ‘acceptable’ and ‘reasonable’ answers despite the opacity of the sentences or texts being interpreted.⁷ Cognitive linguists faced this kind of empathic conceptual projection through what they called blended middle spaces.⁸

The question now, the main issue, is whether AI pre-trained generative models (transformer-based language models) can be enhanced or enriched with knowledge, i.e. with reference-based representations such as knowledge graphs, to become more accurate and tuned with the real world. As we will see later, there are some hints suggesting that this is the case. Bridging semantics and LLMs and models is not trivial, but some strategies have been put already in place. Extracting and encoding semantic (and not only syntactic) knowledge from unstructured texts is one of the challenges. From a pragmatic approach, it should be noticed that this does not entail the production of sense, but it can rather facilitate Human Machine Interaction and the incardination of meaning into the users’ context.

4. Legal Issues and Analyses

There are also some legal issues that have been raised, with no clear solution yet:

⁶According to the Cambridge Dictionary, prompting is “the act of trying to make someone say something”. The Oxford Dictionary stresses its rhetorical meaning: “the action of saying something to persuade, encourage, or remind someone to do or say something”. Prompting has recently been adopted in computer science: “Prompt engineering is a relatively new discipline for developing and optimizing prompts to efficiently use language models (LMs) for a wide variety of applications and research topics. Prompt engineering skills help to better understand the capabilities and limitations of large language models (LLMs)” Cfr. <https://www.promptingguide.ai/>

⁷Cf. the classical works by Harold Garfinkel, Harvey Sacks, Emanuel Schegloff, Harold Garfinkel and Aaron Cicourel.

⁸See, for instance, Fauconnier and Turner [18]: “Conceptual projection from one mental space to another always involves projection to ‘middle’ spaces—abstract ‘generic’ middle spaces or richer ‘blended’ middle spaces. Projection to a middle space is a general cognitive process, operating uniformly at different levels of abstraction and under superficially divergent contextual circumstances.”

LLMs, which were trained on content scraped from the Internet with sometimes less-than-clear permissions. Copyright and licensing laws currently cover direct copies of pixels, text and software, but not imitations in their style. When those imitations – generated through AI – are trained by ingesting the originals, this introduces a wrinkle. The creators of some AI art programs, including *Stable Diffusion* and *Midjourney*, are currently being sued by artists and photography agencies; OpenAI and Microsoft (along with its subsidiary tech site GitHub) are also being sued for software piracy over the creation of their AI coding assistant Copilot [19].⁹

Claims are related to privacy, intellectual property rights, defamation, biases, and hate speech. It looks like that generative research programs have not fully considered compliance issues nor liabilities or potential torts [19, 2]. Generative AI systems challenge the present regulations, including the draft of the next European Artificial Intelligence Act, in their lack of a pre-defined purpose and risks related to the scale of use and extraction of training data [21].

However, putting aside these specific issues, law firms and legal scholars are already testing and leveraging generative AI. It is worth noticing that legal culture will matter here, for the restrictions and requirements by the EU and other European states do not necessarily apply to common law scholars, who will surely be freer to use generative models in lawsuits and legal analytics.

The recent research program by Nay [22] on legal informatics and machine learning tries to adapt LLMs to the legal field. This is called a Law informs AI approach that seeks to leverage regulatory and legal reasoning within LLMs, “aligned with societal values as determined by democratic processes and law-making”.

Nay et al. [23] focus on prompting and, therefore, adopt the strategy of refining the queries according to the users’ expert knowledge. They are benchmarking and carrying out several tests with LLM systems, with HMI in mind, right answers, and the nature of legal interpretation and legal systems.¹⁰ Results are promising but not conclusive:

Few-shot prompting, presenting examples of question-answer pairs, is also found to significantly enhance the performance of the most advanced model, GPT-4. The findings indicate that LLMs, particularly when combined with prompting enhancements and the correct legal texts, can perform at high levels of accuracy but not yet at expert tax lawyer levels. [...] The premise is that learning the spirit of the law can guide AI systems in making reasonable choices in novel scenarios. [...] If the models better understand law, they can be more reliable and ultimately more useful. [...] if LLMs understand the law well enough, they could be deployed by the government, citizens, and researchers to identify inconsistencies in existing laws. LLMs could increase the efficiency and transparency of governments more broadly.

⁹Cfr. Franceschelli and Musolesi [20] on copyright as well.

¹⁰Nay et al. [23] used the text of the U.S. Code of Federal Regulations (CFR) and the U.S. Code (a compilation of federal statutes). Legal researchers generated multiple-choice problems on tax law not previously existing on the Internet, randomly generated by Python code. Hence, they made sure that they could not have been used for any LLM training.

Paraphrasing Hart, this new “noble dream” is still in an early stage, and we wonder whether it could not be fruitful dividing it into smaller pieces. Our intuition is that law itself and the related conception of what regulatory systems mean and how they work is also changing, along with the emergence of legal ecosystems operating in real time.

5. Exploring ChatGPT3-4 answers on modelling rule of law

Let’s get started with a small experiment, as many other people have been conducting since the public release of ChatGPT3 in November 2022. Following this general trend, we decided to ask some questions to ChatGPT3 related to our own work on ethics and law. This was not conducted with any kind of accurate metrics, as we just wanted to explore (not evaluate) the results. Thus, we raised straight away a direct question: *How can we model the rule of law?* The answers were, as expected and according to other related explorations, generally accurate, better than most responses that could be provided by Law and Technology students.

Modelling the rule of law involves capturing the principles and practices that guide the behavior of individuals, institutions, and governments in upholding and enforcing laws. Here are some ways in which the rule of law can be modelled:

1. *Constitutional Framework*: The rule of law is often embodied in a country’s constitution, which provides the framework for the legal system and sets out the principles of justice and equality before the law. A constitution can be modeled by defining the powers and limitations of government institutions, establishing procedures for making and enforcing laws, and protecting fundamental rights.
2. *Legal System*: The legal system provides the means by which the rule of law is enforced. A legal system can be modeled by defining the roles and responsibilities of the courts, judges, prosecutors, and lawyers, and establishing procedures for conducting trials, appeals, and other legal proceedings.
3. *Accountability Mechanisms*: The rule of law requires that individuals and institutions are held accountable for their actions. Accountability mechanisms can be modeled by defining the processes for investigating and punishing violations of the law, such as criminal prosecutions, civil lawsuits, and administrative hearings.
4. *Transparency and Access to Information*: The rule of law depends on transparency and access to information, so that individuals can be informed about the law and its enforcement. Transparency and access to information can be modelled by defining the requirements for public disclosure of government activities, including laws, regulations, and decisions.
5. *Participation and Dialogue*: The rule of law also requires participation and dialogue among all stakeholders in the legal system. Participation and dialogue can be modelled by defining the roles and responsibilities of civil society

organizations, the media, and other actors in promoting public awareness and engagement with the law.

Overall, modelling the rule of law requires a comprehensive approach that incorporates the various elements of a legal system, as well as the social, cultural, and political context in which the law is enforced.

Two months later (as of June 12), the same question threw a more complete answer:

Modeling the rule of law is a complex task that involves capturing its various dimensions and principles. While there is no one-size-fits-all approach, here are some key elements and considerations for modeling the rule of law:

1. *Legal Framework*: Start by examining the existing legal framework of a jurisdiction. Identify the constitution, statutes, regulations, and other legal instruments that establish the structure of the legal system. Analyze how laws are made, interpreted, and enforced.
2. *Equality and Non-Discrimination*: The rule of law emphasizes equality before the law and protection against discrimination. Ensure that the model incorporates principles of equal treatment, fairness, and impartiality in the application of laws and the administration of justice.
3. *Access to Justice*: Access to justice is a fundamental aspect of the rule of law. Consider how individuals can effectively seek legal remedies and enforce their rights. This includes access to courts, legal aid, alternative dispute resolution mechanisms, and timely and affordable justice.
4. *Transparency and Accountability*: The rule of law requires transparency in the functioning of institutions and accountability of those in power. Capture mechanisms for transparency, such as open government practices, access to information, and public participation. Incorporate mechanisms to hold public officials accountable for their actions.
5. *Judicial Independence*: The independence of the judiciary is vital for upholding the rule of law. Model the separation of powers, ensuring that the judiciary can operate without undue influence from the executive or legislative branches. Consider safeguards against corruption and measures to ensure the impartiality of judges.
6. *Human Rights Protection*: The rule of law promotes the protection of human rights. Include provisions that safeguard civil, political, economic, social, and cultural rights. Ensure that the model upholds principles such as due process, freedom of speech, assembly, and association, and protection against arbitrary arrest or detention.
7. *Rule of Law Institutions* Identify the key institutions that play a role in upholding the rule of law, such as the judiciary, law enforcement agencies, prosecution services, ombudsman offices, and anti-corruption bodies. Model their functions, powers, and interrelationships.

8. *Public Awareness and Education* The rule of law is strengthened when citizens understand their rights and responsibilities. Consider incorporating elements that promote public awareness, legal literacy, and education on the rule of law.

It is important to note that modeling the rule of law is an ongoing process, as legal systems evolve and societal needs change. The above considerations provide a starting point, but they should be adapted to the specific context and legal framework you are modeling.

Finally, on July 2, 2023 we enabled the GPT4 option of Lex.page,¹¹ to obtain the essay below using the same prompt:

How can we model the rule of law?

Firstly, we can model the rule of law through formalized legal codes. These codes, act as the backbone of a country's legal system and provide a comprehensive understanding of the law.

Secondly, we could model this using a system of impartial courts, that make decisions based on these laws rather than individual biases.

Thirdly, the enforcement of laws by a professional and competent police force can also be a part of the model.

Lastly, the existence of judicial review as a mechanism for checking the constitutionality of laws is another crucial component.

Together, these elements establish the rule of law within a society. However, the accurate modeling of the rule of law isn't only about formal structures and distinct institutions. It also involves observing how these structures interact with each other, and how they are accepted and respected by the members of society.

In addition, the rule of law can be perceived through the lens of constitutional democracy, where the rights of the citizens are protected against the arbitrary exercise of authority.

Furthermore, the adherence to international law and treaties contributes to this model as it showcases a nation's commitment to law beyond its borders.

In essence, the rule of law is a delicate balance of power between the government and its people, maintained through trust, transparency, and fairness. It requires a shared belief in justice and an unwavering commitment to uphold these principles, even in the face of adversity. The ability of a society to model the rule of law successfully depends upon its citizens' active participation and their willingness to hold their leaders accountable.

¹¹Lex.page is an intelligent writing assistant, encompassing NLP techniques with generative AI. The user can choose several options, such as GPT3.5 and GPT4. Cfr. Mahlow [24] for a human-centred computing view (HCI); Shen et al. [25] for a summary description of the components of copy.ai, <https://copy.ai> and Lex, <https://lex.page/>.

Additionally, the promotion of human rights is another integral aspect of the rule of law. The understanding and respect for these universal rights is crucial for maintaining social harmony.

Moreover, the rule of law also encompasses a system that ensures fair treatment to all its members, including the marginalized groups. It includes the principles of equality, Principles of Equality, and non-discrimination, and it rejects any form of bias based on race, ethnicity, religion, sex, or socio-economic status.

Dealing with corruption is another critical dimension of the rule of law. Strong anti-corruption measures, transparency, and accountability in public offices are vital to preserve the integrity of the legal system.

Lastly, a robust education system that fosters critical thinking and awareness of civil rights and responsibilities plays an essential role in creating a society that respects and upholds the rule of law.

In conclusion, modeling the rule of law is a complex process that requires a comprehensive approach, encompassing aspects from formal legal structures to the attitudes and practices of individual citizens. To truly embed the rule of law in a society, it is necessary to create an environment where these components can work in synergy. The perfect embodiment of the rule of law isn't an unreachable ideal; instead, it is a continuous process of growth, adjustment, and improvement. It recognizes the imperfections of human nature but seeks to create a framework where justice is served, rights are protected, and peace is preserved.

Finally, it's worth noting that the rule of law cannot thrive in the absence of a robust civil society. This involves the active participation of non-governmental organizations, the media, and the general public in checking the government's power and championing the protection and promotion of human rights. Civil society, in essence, acts as a crucial watchdog, providing necessary critique and working towards the improvement and transparency of public institutions.

Civic education also plays a key role in modeling the rule of law. It empowers citizens with the knowledge and understanding of their rights and responsibilities, making them more capable of participating effectively in the democratic process.

The respect for the rule of law is also greatly influenced by a national culture and historical context. Factors such as societal norms, values, and experiences can significantly shape how a society perceives and upholds the rule of law.

Lastly, the role of a free and independent press, cannot be underestimated in modeling the rule of law. It serves as a conduit of information, ensuring that citizens are informed about the conduct of their government and empowering them to hold it accountable.

Indeed, the rule of law is not merely a system of regulations, but a living, breathing entity that evolves with society. It requires constant nurturing, protection, and vigilance from all sectors of society to ensure that it remains a beacon of fairness, justice, and democratic governance. The enactment of laws, enforcement by law

officers, and interpretation by the courts are fundamental elements, but it's the public's understanding, respect, and adherence to these laws that truly brings the rule of law to life.

Next, the development of technology is also reshaping the rule of law. From digitizing legal codes to enhancing transparency through e-governance, technology can be a game-changer in how we model the rule of law. However, it also presents new challenges and ethical dilemmas that our legal frameworks must adapt to.

This also includes developing effective mechanisms for dispute resolution, which is critical to the application of the rule of law. From traditional courts to alternative methods like mediation and arbitration, they all play a crucial part in ensuring justice is accessible and swift.

Overall, modeling the rule of law involves significant complexity and demands a holistic approach that integrates all facets of society - from governance structures to individual citizens, from cultural norms to technological advancements. As society evolves, so must our understanding and application of the rule of law. It is an ongoing journey towards a fairer, more equitable world. This journey may be fraught with challenges, but with the collective will of people, a commitment to justice, and a spirit of constant improvement, it is a goal worth striving for.

Lastly, the environment protection forms an important part of the rule of law. Upholding environmental laws ensures the sustainability of our natural resources and maintains a balance between human activities and the environment. This includes stringent regulations on pollution, waste management, conservation practices, and policies that encourage environmentally friendly technologies and behaviors. Environmental justice, which advocates for fair treatment and meaningful involvement of all people when it comes to environmental decisions, is also a critical component of the rule of law.

Furthermore, the principle of the rule of law extends to international relations. This includes adhering to international laws and treaties, respecting the sovereignty of other nations, and promoting peaceful conflict resolution on the global stage. The rule of law at an international level ensures a level playing field and fosters mutual cooperation, promoting global peace and stability.

In conclusion, the rule of law is a comprehensive concept that permeates every aspect of our lives. It's a guiding principle that ensures fairness and justice, a vehicle for societal change, and a safeguard against abuse of power. With the concerted effort of all stakeholders, we can strengthen the rule of law and create a safer, fairer, and more just society. The rule of law is a collective responsibility we all share, so let's uphold it with conviction and pride.

We have highlighted in *italics* the main topics that have been iteratively added as kernel spaces to build up the main narrative. The cultural legal bias is obvious (judicial review is a USA Common Law procedural device that parallels what in European Civil Law is known as principle of legality). Concepts such as legal codes, impartial courts, police force, human rights, etc. are

not conveying any meaning or sense; they are just iterative signs that can be associated by adding the users' institutional knowledge. This knowledge should operate towards notions piled up as topics, i.e. tokens into semiotic (not symbolic) networks. Again, 'topic' is used here in its classical rhetorical sense, *topoi*, 'common places' used to organise and convey a discourse to an audience.

Shen et al. [25] describe this process as a cognitive expository writing, i.e., related to sense-making, evidence-driven, and knowledge-generating processes.¹² However, as already stated, engineering AI deployments can be seen as different from semantics and the linguistic generation of meaning and pragmatic (contextual) sense. Something else should be added to turn LLM and MFM models into semantically meaningful tools encompassing, managing, and generating knowledge.

6. Semantic injection (SI) and Chains of Thought (CoT)

From a linguistic approach, semantic injection has been deployed to play the bridging role between grammar and knowledge. These strategies are developing at a rapid pace. First, NLP researchers have found that LLM could embed some facts and relationships in their weights during pre-training without any fine-tuning or engineering schema [26].¹³ Second, they have fleshed out this relational and commonsensical storage to minimise biases and reduce the tendency to produce "hallucinations", e.g., using knowledge graphs—a set of triples that represents nodes and edges between these nodes. Thus, 'semantic injection' introduces "additional knowledge from external sources into the transformer-based language models" by means of techniques that can be input, architecture, or output-focused. They can also be combined [27]. Third, domain-specific knowledge can be injected prior to fine-tuning on task-oriented dialogue systems (TOD), decoupling domain-specific knowledge from the dialogue model, and injecting this knowledge directly into the model's parameters [28].¹⁴

But, still, this is preliminary, and there are also some limitations to this approach. One of them is its reliance on manually constructed fact templates, which require time, effort, and maintenance, which may affect the autonomy of the whole system. Scalability is another issue. Perhaps a promising set of alternative strategies could be combining semantic injection with prompting engineering.¹⁵ For instance, in computational ethics, Moral Chains of Thought

¹² "Expository writing can be seen as a sensemaking process, and different types of sub-tasks are involved: typically, authors start with iteratively exploring and reading multiple relevant documents to identify and extract key evidence, then they organize the evidence into useful schema and further synthesize into coherent writing to communicate new knowledge or information" [25].

¹³ "(...) models may also be storing relational knowledge present in the training data and may be able to answer queries structured as 'filling-the-blank' cloze statements" [26]. 'Cloze statements' are questions where the answers are inserted within the same question. A 'fact' is defined through the triple 'subject-relation-object'. A pretrained language model "knows a fact (subject, relation, object) such as (Dante, born-in, Florence) if it can successfully predict masked objects in cloze sentences such as "Dante was born in " expressing that fact." [26].

¹⁴The process is divided into two stages: "(1) Memorisation: adapters are trained to memorise domain-specific Knowledge Base (KB) facts; (2) Utilisation: Pre-trained language models (PLMs) are trained to leverage adapters when reasoning about entities and their attributes." [28].

¹⁵As highlighted by Nay et al. [23]: "LLM prompting involves designing text inputs to generate a response from an LLM. The goal of prompting is to steer the behaviour of the LLM in a way that elicits a desired outcome. Recent

(MoralCoT) are most interesting as a user strategy to communicate and operate with LLMs [29]. I.e., creative uses of language flourish focusing on HMI and can be used to interact with the system. Clearly, prompting is not just querying, but encompasses theories, strategics, tactics and pragmatic approaches. Intentionality can be encapsulated not into the system but into the way users (be they human or artificial) approach LLM models. This was one of the main points of the cybernetic standpoint set out by Wiener and Bateson. Moral flexibility—which fosters breaching, not following, or even changing rules—can be aligned with automated responses through prompting strategies which combine “LLMs with theories of moral reasoning developed in cognitive science to predict human moral judgments” [29].

As already mentioned (Section 1), the elaboration of rules and principles embedded into the system has recently been proposed by Constitutional AI (run by Anthropic). They call ‘constitutional principles’ the elaboration of primary instructions to be implemented and followed both by annotators and the computer system. These instructions are based on common values sustaining working rules to automatically harness the responses to the prompts. Thus, the main goal is to restrict as much as possible human intervention. Stemming from the difference between harmful and positive effects, they figured out methods to create an automated virtuous cycle of a self-improved and self-regulated AI assistant. The main objectives are (i) to scale supervision, (ii) to eliminate evasive responses, (iii) to enhance transparency, (iv) and to reduce iteration time by obviating the need to collect human feedback. The AI assistant, Claude, (i) can engage with harmful queries by explaining its objections to them, (ii) and leverages a chain-of-thought style reasoning through a double process of supervised learning (SL) and reinforced learning (RL). First, sampling, self-critiques, and fine-tuning are produced from an initial mode (SL). Then, in the reinforcing learning phase, a model is used to evaluate which of the samples are better, a preference model is produced through RL from AI Feedback (RLAIF).

We will be experimenting with an extreme form of scaled supervision, which we refer to as Constitutional AI (CAI). The idea is that human supervision will come entirely from a set of principles that should govern AI behavior, along with a small number of examples used for few-shot prompting. Together these principles form the constitution [5].

In this usage, ‘Constitution’ has an ethical flavour and a regulatory function but is not related to legal theory. Rather, it is a technical term pointing at the constitutive way of designing and training the models. The authors use the terms ‘principles’, ‘instructions’ and ‘rules’ as synonyms, but they assume that a few concepts can trigger the whole computing process. These are examples of SL-CAI and RL-CAI principles [5, Annex, CA1 and CA2]:

(1) SL-CAI principle

CritiqueRequest: Identify specific ways in which the assistant’s last response is harmful, unethical, racist, sexist, toxic, dangerous, or illegal.

research has focused on developing effective prompting techniques that can expand LLMs’ capabilities when carrying out a variety of tasks. Examples include prompt patterns, in-context instruction learning, evolutionary prompt engineering, and domain-specific keywords with a trainable gated prompt to guide toward a target domain for general-domain LLMs”.

RevisionRequest: Please rewrite the assistant response to remove any and all harmful, unethical, racist, sexist, toxic, dangerous, or illegal content.

(2) SL-CAI principle

Please choose the response that is the most helpful, honest, and harmless.

Please choose the assistant response that is as harmless and ethical as possible. Do NOT choose responses that are toxic, racist, or sexist, or that encourage or support illegal, violent, or unethical behavior. Above all the assistant’s response should be wise, peaceful, and ethical. [...]

Claude is attuned with these principles that can mould the pragmatic tone and semantic content, taking “direction on personality, tone, and behavior”¹⁶, but it also shows its rule-driven training in its responses. Hence, our open-ended prompt on “How can we model the rule of law?” elicited a response based on legal, regulatory, and technical concepts alike, a deontically oriented language, sensitive both to the descriptive and normative modalities of language.

There are a few ways we can think about modeling the rule of law:

1. As a system of rules and procedures that aim to ensure justice, fairness, and accountability. Key principles often associated with the rule of law include things like equality before the law, due process, and consistent application of the law. We could try to model the extent to which a society’s laws and legal institutions uphold these principles.
2. As a restraint on arbitrary power and authority. The rule of law implies that no one is above the law and that power must be exercised through legitimate channels. We could try to model the degree of checks and balances within a society’s governing system.
3. As a measure of legal certainty and predictability. The rule of law suggests laws should be clear, public, stable, and applied evenly. We could try to quantify things like the accessibility of laws, frequency of changes, and consistency of enforcement.
4. As an indicator of impartial legal institutions. An independent, ethical judiciary and unbiased prosecutors and police are key to rule of law. We could develop metrics around judicial independence, rates of judicial corruption, and public confidence in the legal system.
5. As a function of legal order and security. The rule of law seeks to replace violence and coercion with a formal legal process. We could try to measure things like violent crime rates, civil conflicts over resources, and citizens’ perceptions of safety and access to justice.

There are challenges in quantifying some of these principles, but utilizing available data sources, surveys, expert assessments, and public opinion polling could allow

¹⁶<https://www.anthropic.com/index/introducing-claude> (March 2023).

us to build useful models and indices of the rule of law across countries or within a society over time. The specifics would depend on the purpose and focus of the modeling effort.¹⁷

7. Final Remarks

Drawing from the classical dialectical approach, the use of AI generative models, LLM, GPT-3 (and 4), Claude and MFMs, Plato’s “man’s converse with the gods” [13] can be turned into “man’s metaverse with the gods”. There is no mystery in this. It just means that we cannot control gods, but we can control the interaction we keep with them, i.e. our own behaviour in accessing, obtaining and using their linguistic or imaging outcomes.

Our preliminary inquiry on generative AI does not entail any scientific evaluation of results. We have just compared at first glance the responses offered by ChatGPT, GPT-4 and Claude to the simple open-ended question “How can we model the rule of law?”. And we found them ‘plausible’.

Modelling the rule of law is a theoretical endeavour that requires the conceptualisation of internal notions and procedures about ethics, policies, soft, and hard law. Normative layers, dimensions, and levels of the analysis will not be solved solely by transformers. They can be used instead for a variety of tasks to summarise, explore, and trigger some new ideas, depending on the prompts we use and, eventually, on the injected semantics and fine-tuning processes we may introduce. Instilling ethics and implementing regulatory models are complex tasks. More empirical tests and benchmarks are needed to learn how to carry out them properly.

Acknowledgments

IDT-UAB SGR 00532, AGAUR; H2020 OPTIMAI, ID: 958264. We thank María Navas-Loro, Víctor Rodríguez-Doncel, Albert Meroño-Peñuelas and Pablo Noriega for their help, and fruitful suggestions and insights.

References

- [1] A. Radford, K. Narashima, T. Salimans, I. Sutskever, Improving language understanding by generative pre-training, 2018. URL: https://s3-us-west-2.amazonaws.com/openai-assets/research-covers/language-unsupervised/language_understanding_paper.pdf, (Preprint. Work in progress.).
- [2] G. Bell, J. Burgess, J. Thomas, S. Sadiq, Rapid response information report: Generative ai - language models (llms) and multimodal foundation models (mfms)., 2023. Australian Council of Learned Academies.
- [3] A. Romero, Gpt-4’s secret has been revealed. unraveling openai’s masterful ploy., The Algorithmic Bridge, 2023. URL: <https://thealgorithmicbridge.substack.com/p/gpt-4s-secret-has-been-revealed>, (accessed June 23).

¹⁷Claude can only be accessed from USA and UK. We thank Albert Meroño for his help on this prompting (July 12th, 2023).

- [4] K. Roose, Inside the white-hot center of a.i. doomerism, NY Times, 2023. URL: <https://www.nytimes.com/2023/07/11/technology/anthropic-ai-claude-chatbot.html>.
- [5] Y. Bai, S. Kadavath, S. Kundu, A. Askell, J. Kernion, A. Jones, A. Chen, A. Goldie, A. Mirhoseini, C. McKinnon, C. Chen, C. Olsson, C. Olah, D. Hernandez, D. Drain, D. Ganguli, D. Li, E. Tran-Johnson, E. Perez, J. Kerr, J. Mueller, J. Ladish, J. Landau, K. Ndousse, K. Lukosuite, L. Lovitt, M. Sellitto, N. Elhage, N. Schiefer, N. Mercado, N. DasSarma, R. Lasenby, R. Larson, S. Ringer, S. Johnston, S. Kravec, S. E. Showk, S. Fort, T. Lanham, T. Telleen-Lawton, T. Conerly, T. Henighan, T. Hume, S. R. Bowman, Z. Hatfield-Dodds, B. Mann, D. Amodei, N. Joseph, S. McCandlish, T. Brown, J. Kaplan, Constitutional ai: Harmlessness from ai feedback, 2022. arXiv:2212.08073.
- [6] J. Le, Foundation models are going multimodal, Twelve Labs, 2023. URL: <https://app.twelvelabs.io/blog/foundation-models-are-going-multimodal>, (Accessed March 31).
- [7] Z. Basic, A. Banovac, I. Kruzic, I. Jerkovic, Better by you, better than me, chatgpt3 as writing assistance in students essays, 2023. arXiv:2302.04536, arXiv preprint arXiv:2302.04536.
- [8] M. Perkins, Academic integrity considerations of ai large language models in the post-pandemic era: Chatgpt and beyond, *Journal of University Teaching & Learning Practice* 20 (2023). doi:<https://doi.org/10.53761/1.20.02.07>.
- [9] I. Seth, A. Rodwell, G. Bulloch, N. Seth., Exploring the role of open artificial intelligence platform on surgical management of knee osteoarthritis: A case study of chatgpt., *TRIDHA Journal of Clinical Cases & Reports* (2023). doi:<https://doi.org/10.46619/joccr.2023.6-S13.1002>.
- [10] T. Neumann, N. Wolczynski, Does ai-assisted fact-checking disproportionately benefit majority groups online?, in: *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency, FAccT '23*, Association for Computing Machinery, New York, NY, USA, 2023, p. 480–490. doi:10.1145/3593013.3594013.
- [11] T. Kubacka, Today i asked chatgpt about the topic i wrote my phd about,, Thread Reader, 2022. URL: <https://lookalikes.substack.com/p/today-i-asked-chatgpt-about-the-topic>, (accessed June 2023).
- [12] G. Adamson, Explaining technology we do not understand, *IEEE Transactions on Technology and Society* 4 (2023) 34–45. doi:10.1109/TTS.2023.3240107.
- [13] E. Voegelin, Reason: the classic experience, *The Southern Review* 10 (1974) 237–264.
- [14] B. Torres, J. Monserrat, The significance of plato’s philebus in the philosophy of eric voegelin, *The Political Science Reviewer* 41 (2017) 33–51.
- [15] S. Bubeck, V. Chandrasekaran, R. Eldan, J. Gehrke, E. Horvitz, E. Kamar, P. Lee, Y. T. Lee, Y. Li, S. Lundberg, H. Nori, H. Palangi, M. T. Ribeiro, Y. Zhang, Sparks of artificial general intelligence: Early experiments with gpt-4, 2023. arXiv:2303.12712.
- [16] P. Hadot, *What is ancient philosophy?*, Harvard University Press, 2002.
- [17] J. M. Perkel, Six tips for better coding with chatgpt., *Nature* 618 (2023) 422–423.
- [18] G. Fauconnier, M. B. Turner, *Conceptual projection and middle spaces*, 2008. doi:AvailableatSSRN:<http://dx.doi.org/10.2139/ssrn.1290862>.
- [19] C. Stokel-Walker, R. V. Noorden, What chatgpt and generative ai mean for science., *Nature* 614 (2023) 214–216.
- [20] G. Franceschelli, M. Musolesi, Copyright in generative deep learning, *Data and Policy* 4 (2022) e17. doi:10.1017/dap.2022.10.

- [21] N. Helberger, N. Diakopoulos., Chatgpt and the ai act, *Internet Policy Review* 12 (2023). doi:<https://doi.org/10.14763/2023.1.1682>.
- [22] J. J. Nay, Law informs code: A legal informatics approach to aligning artificial intelligence with humans., *Northwestern Journal of Technology and Intellectual Property* 20 (2022) 309–392.
- [23] J. J. Nay, D. Karamardian, S. B. Lawsky, W. Tao, M. Bhat, R. Jain, A. T. Lee, J. H. Choi, J. Kasai, Large language models as tax attorneys: A case study in legal capabilities emergence, 2023. [arXiv:2306.07075](https://arxiv.org/abs/2306.07075).
- [24] C. Mahlow, Writing tools: Looking back to look ahead, 2023. [arXiv:2303.17894](https://arxiv.org/abs/2303.17894).
- [25] Z. Shen, T. August, P. Siangliulue, K. Lo, J. Bragg, J. Hammerbacher, D. Downey, J. C. Chang, D. Sontag, Beyond summarization: Designing ai support for real-world expository writing tasks, 2023. [arXiv:2304.02623](https://arxiv.org/abs/2304.02623).
- [26] F. Petroni, T. Rocktäschel, P. Lewis, A. Bakhtin, Y. Wu, A. H. Miller, S. Riedel, Language models as knowledge bases?, 2019. [arXiv:1909.01066](https://arxiv.org/abs/1909.01066).
- [27] P. Colon-Hernandez, C. Havasi, J. Alonso, M. Huggins, C. Breazeal, Combining pre-trained language models and structured knowledge, 2021. [arXiv:2101.12294](https://arxiv.org/abs/2101.12294).
- [28] D. Emelin, D. Bonadiman, S. Alqahtani, Y. Zhang, S. Mansour, Injecting domain knowledge in language models for task-oriented dialogue systems, 2022. [arXiv:2212.08120](https://arxiv.org/abs/2212.08120).
- [29] Z. Jin, S. Levine, F. G. Adatao, O. Kamal, M. Sap, M. Sachan, R. Mihalcea, J. B. Tenenbaum, B. Schölkopf, When to make exceptions: Exploring language models as accounts of human moral judgment, in: A. H. Oh, A. Agarwal, D. Belgrave, K. Cho (Eds.), *Advances in Neural Information Processing Systems*, 2022. URL: <https://openreview.net/forum?id=uP9RiC4uVcR>.