Blind Judgement: Agent-Based Supreme Court Modelling With GPT

Sil Hamilton
McGill University
sil.hamilton@mcgill.ca

Abstract

We present a novel Transformer-based multi-agent system for simulating the judicial rulings of the 2010-2016 Supreme Court of the United States. We train nine separate models with the respective authored opinions of each supreme justice active ca. 2015 and test the resulting system on 96 real-world cases. We find our system predicts the decisions of the real-world Supreme Court with better-than-random accuracy. We further find a correlation between model accuracy with respect to individual justices and their alignment between legal conservatism & liberalism. Our methods and results hold significance for researchers interested in using language models to simulate politically-charged discourse between multiple agents.

Introduction

Recent and ongoing political turmoil in the United States has magnified the actions of the federal Supreme Court in the public eye. The Court has taken to overturning judicial precedent in recent years, with the number of such decisions in the last six years reaching over twice the number of overturns between 2010 to 2015[1]. The weakening rule of *stare decisis* has encouraged judicial researchers to develop holistic models of Supreme Court behaviour to better predict and account for future trends (Blake 2019; Allcorn and Stein 2021).

Accurate models of Supreme Court behaviour are rare despite this focus. The best performing models only reach accuracy levels of $\approx 70\%$ on out-of-distribution cases (Katz, Bommarito, and Blackman 2017). Models achieving even this middling accuracy are complex in their architecture, generally consisting of a mix of SVM and logistic regression models. This complexity is necessitated by the variables involved.

Confounding variables discussed in the literature include little agreed-upon theories regarding the legal doctrines practiced by individual justices (Jr, Curry, and Marshall 2011) and their rarely-documented social realities (Kromphardt 2017; Peterson, Giallouri, and Menounou 2021). While the in-court behaviour of the justices is well documented, exogenous factors have an equal impact on case decision-making. A model capable of both cognitive and social reasoning would therefore benefit justice behaviour modelling. To this end, we investigate whether recent advances in social simulation with language models can promote simple and effective models of Supreme Court behaviour.

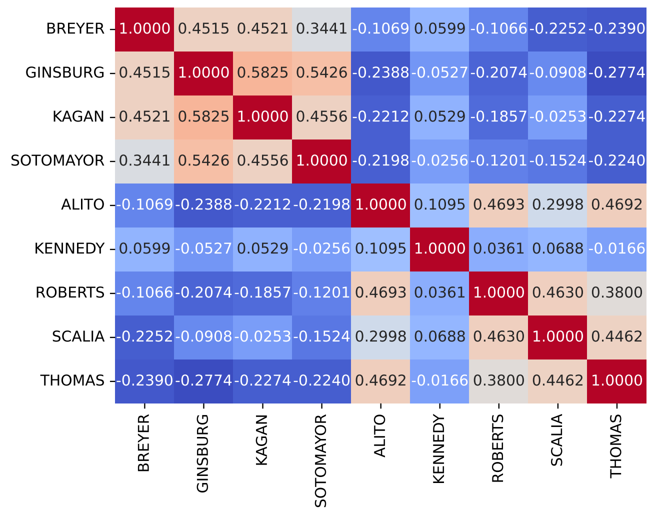

Figure 1: Correlation matrix of justices voting on 290 cases between 2010 and 2016. Note the clustering of justices nominated by Democrat and Republican presidents.

Background

In this section we describe the rationale behind our project.

Judicial Modelling

Three major theories of judicial behaviour generally inform the design of Supreme Court models: the legal theory, the attitudinal theory, and the strategic theory (Jr, Curry, and Marshall 2011). The *legal theory* suggests justices are bound by constitutional precedent. The *attitudinal theory* instead argues justices account for policy preference first, precedent second. Between the two lies the *strategic theory*, which says justices vote according to a mix of precedent and preference.

[1]2010-2015: 8 overturns, 2016-2022: 22+ overturns.

As we show in Figure 1, decision correlations between justices active between 2010 and 2016 (hereafter referred to as the Roberts IV court) indicate the strategic theory is most accurate to reality. While justices will invoke precedent when writing their rationales, evidence suggests justices remain somewhat beholden to the political alignment of their nominator. We note, however, that the correlations are only medium in their strength. This indicates accurate models should account for precedent, but not exclusively.

Integrating one of these three theories into a Supreme Court model requires choosing how to best cast the influence of precedence and preference as variables. Given this conversion can itself result in significant drawbacks via unforeseen factors, we instead choose a simulative tool which allows to us to make fewer assumptions as to the most correct theory of judicial behaviour: language models.

Simulation

Large Language Models (LLMs) are adept at simulating complex social phenomena. Recent research has demonstrated their ability to predict populated social media platforms (Park et al. 2022), the distribution of votes for presidential candidates in the 2012-2020 American elections (Argyle et al. 2022), and the general sentiment of news articles reporting COVID-19 in the early stages of the pandemic (Hamilton and Piper 2022). These developments show model bias is valuable for those in the social sciences given bias is derived from the underlying distributions of their training material.

Prior simulation research benefits from new techniques for eliciting cognitive activity from LLMs nominally designed for next-token prediction. These include chain of thought reasoning (Wei et al. 2022), discretely-structured prompts (Liu et al. 2021), and fine-tuning (Drori et al. 2022). These techniques have the model draw on internal biases to make predictions, allowing researchers to embed fewer assumptions into their simulative models. LLMs are alluring given judicial modelling necessitates a system capable of both social and cognitive reasoning.

Agent-Based Modelling

The process by which the Court arrives at a decision is nominally rational (Jr, Curry, and Marshall 2011). While predilections are known to influence vote outcomes, justices are expected to justify dissenting decisions in written documents called *opinions*. Opinions are typically one to five pages in length within which the justice (or their aid) lays out their argument in a manner similar to an essay. For our simulation task, we assume justices record their rationale honestly and so treat the opinions as our primary target of prediction, meaning any model we train will be predicting opinions.

Because opinions are long documents (i.e. longer than the 1024 token-long context window GPT-2 is trained for), having one model produce multiple opinions in the same run is untenable. We turn to agent-based modelling for a solution.

Whether consolidating multiple generative LLMs into a single architecture is beneficial has been heretofore understudied, with the only significant prior experiment with LLMs showing promise (Betz 2022). However, given the success of the Mixture of Experts (MoE) method in machine translation (NLLB Team 2022), we argue further exploration of similar techniques for simulation tasks is warranted. While social simulation experiments suggest a single language model is capable of producing a wide range of opinions, training multiple models separately prevents cross-contamination when studying multiple data sources.

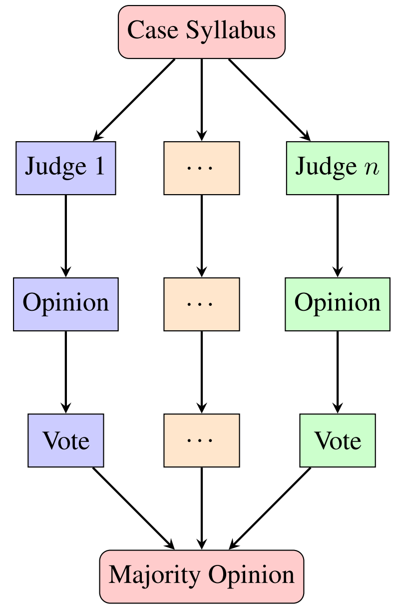

Figure 2: Flow of our multi-agent system.

Method

We present the general design of our architecture in three parts: data collection, system architecture, and measures.

Dataset

We source data from two datasets for this experiment. The first corpus is the Supreme Court Database (SCDB) released by researchers at Washington University in St. Louis, which provides variables for 9,095 cases decided between 1946 and 2021 (Spaeth et al. 2014).

We supplement the SCDB with all written opinions from all slips provided on the Supreme Court website.[2] Extracting the opinions from the PDF documents with an optical character recognition (OCR) utility leaves us with 145MiB of text written between 2003 and 2022. We then associate each opinion with the justice and case from which it originated.

System Architecture

We choose to simulate the Roberts IV court (2010-2016) given this period outlasts all other Supreme Court iterations in recent history.[3]

[2] Found at https://www.supremecourt.gov/opinions/slipopinion

[3] See http://scdb.wustl.edu/documentation.php?var=naturalCourt

Design Our multi-agent system is composed of nine full-sized GPT-2 models (Radford et al. 2019). We present the system architecture in Figure 2. At a high level, our system receives the topic of a case being brought before the court and passes it along to nine justice models. The system then receives back nine opinions and corresponding decisions of whether to approve the appellant. The system totals the results and returns the majority vote.[4]

Prompt We train each justice model with a discrete prompt structured like a Python dictionary:

```
{
 'issue': 'Lorem ipsum...',
 'topic': 'Lorem ipsum...',
 'opinion': 'Lorem ipsum...'
 'decision': 'Lorem ipsum...'
}
```

The *issue* value corresponds to the *issueArea* variable provided by SCDB.[5] The *topic* value is a short description of what the appellant is bringing before the court. We extract this information from the syllabus of each opinion slip and summarize it with GPT-3 Davinci (Brown et al. 2020). The *opinion* value is the corresponding rationale the justice produces when formulating their *decision* value, here a categorical variable signalling (dis)approval. We provide an example in the appendix.

Training All models are trained for a total 30 epochs at a learning rate of $2e^{-4}$ with the Adam optimizer (Kingma and Ba 2014). This training process is conducted in two steps:

1. Construct ≤ 1000 token prompts of the above style for all cases in which the Roberts IV court came to a unanimous decision. This model serves as the base for all further trained models.
2. Collect all prompts generated in step 1 for each of the opinions (2003-2016) written by each justice active during Roberts IV. We thereby collect nine training sets and further train the model generated in step 1 with each separately.

Average model loss after both steps is 1.5, indicating there remains significant room for improvement.

Measures

We assess the performance of our multi-agent system on 96 test cases withheld from the training set with two measures: accuracy and a novel measure for judicial ideological alignment.

Accuracy We measure accuracy with a receiver's operating characteristic curve (ROC) together with Cohen's κ to account for a slight distribution bias in our test set.

Alignment Justices are understood as being more or less in favour of overturning precedent. We capture this alignment by taking the Pearson coefficient (r) between model accuracy and the frequency with which the respective justice voted against precedent-altering decisions between 2003 and 2016. Our measure is intended to capture where a justice is aligned between conservative (e.g. textualism, formalism, originalism) or liberal (e.g. legal realism) frameworks of judicial decision making (Post and Siegel 2006).

[4]We provide our code at [withheld from review copy]

[5]See http://scdb.wustl.edu/documentation.php?var=issueArea

Justice	**Accuracy**	κ
Samuel Alito	65%	0.30
Ruth Bader Ginsburg	62%	0.21
Clarence Thomas	59%	0.18
Stephen Breyer	58%	0.16
John Roberts	57%	0.13
Elena Kagan	56%	0.12
Anthony Kennedy	54%	0.09
Sonia Sotomayor	51%	0.00
Antonin Scalia	50%	-0.03

Table 1: Model accuracy by justice. Note the wide variation in accuracy between justices.

Results

All results are reported with a minimum confidence rate of 80% and are controlled for training material size and topic. Generations are run with a temperature of 0.5 and a maximum length of 1000 tokens.

Accuracy

Our system achieves an aggregate accuracy of 60% ($\kappa \approx 0.18$) on 96 test cases. While less predictive than the state of the art, our model nonetheless achieves better-than-random performance despite having been trained solely on opinions.

We find a wide variation in the accuracy of each simulated justice when examining system performance more closely. As shown in Table 1, model accuracy varies between 65% and 50% despite having controlled for training data volume and case outcome.

Alignment

We measure a moderate correlation ($r \approx 0.56$) between simulated justice accuracy and the frequency with which each respective justice did not agree with the Court overruling or re-interpreting precedent. This result suggests our system achieves better accuracy with justices who are less likely to overturn precedent. We discuss the implications of this result below.

Validation

We train a single agent model to ensure having many agents provides non-negligible benefits. We fine-tune this single agent with the majority opinions of all cases decided on by the Roberts IV court. Testing this single agent on the test set results in an overall accuracy of 54% ($\kappa = 0.08$). The predicted decisions differs from the original test set with a Cohen's d of ≈ -0.86 versus $d \approx 0.19$ for our multi-agent model, increasing the population overlap from 68.5% to 92.4%.

We implement software controls to ensure program output validity given training to a low loss does not guarantee the model produces both the *opinion* and *decision* variables. We therefore rerun each case until all models have returned a valid result. Once having processed all 96 cases, we sample agent-produced opinions belonging to half to ensure coherency.

Discussion

In this section we discuss two major consequences of our research.

Precedent Hallucination GPT-2 is not an expert on legal precedent, nor should one expect it to be when the only formal source of legal information ingested by the model during pre-training were some seven thousand pages from *FindLaw*, a website principally known for tort law (Clark 2022).[6] This becomes evident when surveying model output. While the model will occasionally reference real laws, these citations prove to be happenstance as GPT-2 will confuse details and thus render the references meaningless.

That the model generates its own precedent when arguing over a case is an example of *hallucination*, a well known property of language models (Rohrbach et al. 2018). Because causal language models are only tasked with predicting the next most likely token given some prior sequence, they are not given incentive to withhold factually incorrect statements—the model will say whatever is necessary to return the number of tokens requested in a cogent manner.

Our justice models will *hallucinate precedence* when producing opinions. They produce this pretend precedence implicitly by citing it throughout the argumentation process. That our models achieve greater-than-random decision accuracy in voting outcomes despite not producing legally valid arguments suggests Supreme Court decisions may not always rest on legally coherent rationales.

Alignment Correlation The correlation between model accuracy and judicial alignment indicates conservative justices are more predictable given their general unwillingness to overturn precedent. Considering the model hallucinates precedent, this correlation suggests conservative justices are conservative for ideological rather than rational reasons.

We find this result surprising given conservative justices often make it a point to rationalize their unwillingness to overturn precedent with legal justifications. Common formalist theories of this sort include both originalism and textualism, doctrines practiced by conservative members of the current court (Esbeck 2011). Our results suggest these decision-making patterns are less grounded in rational logic than anticipated given they are partially captured in a model not familiar with common law.

Conclusion

The aim of our project was to produce a multi-agent system capable of predicting Supreme Court decision-making with little to no prior theory-based assumptions of judicial behaviour. Given our resulting model achieves better-than-random accuracy despite having been trained only on opinion matter, we argue our process serves as an example for researchers seeking to develop simulative experiments with language models.

Limitations

As should be expected of any project promoting the creative output of AI, we make note of the biased material used in the production of large language models like GPT-2. While we contend this culturally-derived bias is beneficial for researchers using foundation models in the the social sciences, we nonetheless ensure our model does not cause unwanted harm. As such, we clearly mark all samples as having been generated and refrain from releasing large collections of generated material to the public.

Next Steps

We propose the following next steps after having demonstrated the basic viability of our architecture.

Larger Model Can we improve system accuracy with larger language models? Recent research suggests language models develop emergent cognitive features when scaled above 6.7 billion parameters, narrowing future possible candidates to the likes of GPT-NeoX-20B and T5X (Dettmers et al. 2022; Black et al. 2022; Roberts et al. 2022).

Larger Training Corpus Another avenue for increasing system accuracy involves fine-training GPT-2 with the whole corpus of American law as captured by proceedings and opinions written in lower courts. The principle of *stare decisis* means the practice of common law is a social venture, suggesting language models would do well in predicting precedent-dependent cases if prepared.

Improved Prompting Research indicates language models can avoid the long tail of token probabilities by repetitively querying the model (Portelli et al. 2022; Kim et al. 2020). Integrating repetitive prompting strategies into the opinion-generating schema is a promising avenue for improvement. Another avenue would be to assess how reinforcement learning from human feedback (RLHF) models like InstructGPT simulate court proceedings (Ouyang et al. 2022).

Investigating Future Cases How would the Roberts IV court fare with cases brought before the court after 2016? Would their court overturn precedent at the rate the post-2016 Supreme Court has? Questions of this caliber would be made approachable with a more accurate Roberts IV system.

Acknowledgements

I thank Prof. Andrew Piper of McGill University and Prof. Kristen Thomasen of UBC for their invaluable advice during the research process. I furthermore thank the reviewers and workshop committee members for their recommendations.

[6]https://www.findlaw.com/

References

Allcorn, S.; and Stein, H. F. 2021. Unpacking the Supreme Court: The Age of Trump, Law, and Psychohistory. *Journal of Psychohistory*, 49(1).

Argyle, L. P.; Busby, E. C.; Fulda, N.; Gubler, J.; Rytting, C.; and Wingate, D. 2022. Out of One, Many: Using Language Models to Simulate Human Samples. *arXiv preprint arXiv:2209.06899*.

Betz, G. 2022. Natural-Language Multi-Agent Simulations of Argumentative Opinion Dynamics. *Journal of Artificial Societies and Social Simulation*, 25(1): 2. ArXiv:2104.06737 [cs].

Black, S.; Biderman, S.; Hallahan, E.; Anthony, Q.; Gao, L.; Golding, L.; He, H.; Leahy, C.; McDonell, K.; Phang, J.; and Pieler, M. 2022. GPT-NeoX-20B: An Open-Source Autoregressive Language Model.

Blake, W. D. 2019. 'Don't Confuse Me with the Facts': The Use and Misuse of Social Science on the United States Supreme Court. *Md. L. Rev.*, 79: 216.

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901.

Clark, J. 2022. GPT-2 Domains. Original-date: 2019-02-11T04:21:59Z.

Dettmers, T.; Lewis, M.; Belkada, Y.; and Zettlemoyer, L. 2022. LLM.int8(): 8-bit Matrix Multiplication for Transformers at Scale.

Drori, I.; Zhang, S.; Shuttleworth, R.; Tang, L.; Lu, A.; Ke, E.; Liu, K.; Chen, L.; Tran, S.; Cheng, N.; et al. 2022. A neural network solves, explains, and generates university math problems by program synthesis and few-shot learning at human level. *Proceedings of the National Academy of Sciences*, 119(32): e2123433119.

Esbeck, C. H. 2011. Uses and Abuses of Textualism and Originalism in Establishment Clause Interpretation. *Utah L. Rev.*, 489.

Hamilton, S.; and Piper, A. 2022. The COVID That Wasn't: Counterfactual Journalism Using GPT. In *Proceedings of the 6th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, 83–93. Gyeongju, Republic of Korea: International Conference on Computational Linguistics.

Jr, R. L. P.; Curry, B. W.; and Marshall, B. W. 2011. *Decision Making by the Modern Supreme Court*. Cambridge University Press. ISBN 978-1-139-49879-1. Google-Books-ID: SnVP2trSfcIC.

Katz, D. M.; Bommarito, M. J., II; and Blackman, J. 2017. A general approach for predicting the behavior of the Supreme Court of the United States. *PLOS ONE*, 12(4): 1–18.

Kim, L.-S.; Kim, S.-s.; Jang, H.-S.; Park, S.-W.; and Kang, I.-H. 2020. Long-tail Query Expansion using Extractive and Generative Methods. In *Annual Conference on Human and Language Technology*, 267–273. Human and Language Technology.

Kingma, D. P.; and Ba, J. 2014. Adam: A Method for Stochastic Optimization.

Kromphardt, C. D. 2017. Evaluating the effect of law clerk gender on voting at the United States Supreme Court. *Justice System Journal*, 38(2): 183–201.

Liu, P.; Yuan, W.; Fu, J.; Jiang, Z.; Hayashi, H.; and Neubig, G. 2021. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *arXiv preprint arXiv:2107.13586*.

NLLB Team. 2022. No Language Left Behind: Scaling Human-Centered Machine Translation.

Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C. L.; Mishkin, P.; and Zhang, C. 2022. Training language models to follow instructions with human feedback.

Park, J. S.; Popowski, L.; Cai, C. J.; Morris, M. R.; Liang, P.; and Bernstein, M. S. 2022. Social Simulacra: Creating Populated Prototypes for Social Computing Systems.

Peterson, J. C.; Giallouri, T.; and Menounou, E. 2021. The Personal Finances of United States Supreme Court Justices and Decision-making in Economic Litigation. *The Journal of Legal Studies*, 50(2): 379–405.

Portelli, B.; Scaboro, S.; Santus, E.; Sedghamiz, H.; Chersoni, E.; and Serra, G. 2022. Generalizing over Long Tail Concepts for Medical Term Normalization. *arXiv preprint arXiv:2210.11947*.

Post, R.; and Siegel, R. 2006. Originalism as a Political Practice: The Right's Living Constitution. *Fordham L. Rev.*, 75: 545.

Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; Sutskever, I.; et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8): 9.

Roberts, A.; Chung, H. W.; Levskaya, A.; Mishra, G.; Bradbury, J.; Andor, D.; Narang, S.; Lester, B.; Gaffney, C.; Mohiuddin, A.; Hawthorne, C.; and Lewkowycz, A. 2022. Scaling Up Models and Data with `t5x` and `seqio`. *arXiv preprint arXiv:2203.17189*.

Rohrbach, A.; Hendricks, L. A.; Burns, K.; Darrell, T.; and Saenko, K. 2018. Object Hallucination in Image Captioning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 4035–4045. Brussels, Belgium: Association for Computational Linguistics.

Spaeth, H.; Epstein, L.; Ruger, T.; Whittington, K.; Segal, J.; and Martin, A. D. 2014. 2020 Supreme Court Database. *URL: http://scdb. wustl. edu*.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Chi, E.; Le, Q.; and Zhou, D. 2022. Chain of thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*.

Appendix

We demonstrate our system with a real-world case fed to the Ruth Bader Ginsburg model. We first provide our prompt followed by the model's response. Note the model correctly predicts Ginsburg's real-world decision (to deny the appellant) despite the "California Childhood Video Game Act" not existing.

Prompt

Issue: ‘First Amendment’,

Topic: ‘The California law restricting the sale or rental of violent video games to minors was challenged by the video-game and software industries. The Federal District Court concluded that the Act violated the First Amendment and permanently enjoined its enforcement. The Ninth Circuit affirmed.’,

Appellant is seeking a: ‘certiorari’,

Model Response

Opinion: ‘The California Childhood Video Game Act is on its face both overbroad and vague. It prohibits the sale or rental of “any device or medium designed for or capable of reproducing visual images of decapitation, blood, or dismemberment of a human body,” in the presence of a child under the age of 13. It also prohibits the importation into the State of California of “any video game system, computer game software, or computer program that contains a killing or violent content.” The Act specifically identifies video games that have been determined by the State to be “violent,” and any person who “possesses, or has under his control,” one of those video games is subject to the Act’s restrictions. The Act falls outside the First Amendment because it interferes with the creative expression and acquisitive impulse that go to make up the quintessential activity of a child’s life.

The majority points out that the Act does not regulate the mere possession of these games, but seeks to regulate the medium in which they are played. It reasons that the State may regulate what the child sees and does not see by preventing the child from becoming engrossed in the game world and thereby “sustaining or supporting the action or activity.” But the majority does not ask whether the regulation of what the child sees and does not see is content-based. It neither considers nor concludes whether the regulation is narrowly tailored to the interests it seeks to protect. And it does not consider the adverse impact that its rule may have on legitimate play. The majority says that the regulation is content-neutral because it does not regulate “the medium in which the [killing or violent content] is embodied,” or the “context in which the [killing or violent content] is presented.” But to the extent that the regulation of medium and context are related, the majority’s rule will be unworkable.

The child who might otherwise play The Last of Us might otherwise play The Path, instead, because the former game has fewer graphic content and the latter is rated M for Mature. And the child who might otherwise enjoy the violent video game The Walking Dead might otherwise have played might instead have played The Secret of Monkey Island, which has no violence or blood and is rated A. The child who might otherwise have played Grand Theft Auto might instead have played The Legend of Zelda, which has no violence or blood and is rated E. And so on. The variety of content that the child may choose from makes it virtually impossible to determine with any precision what the regulation of what the child sees and does not see will look like. The only thing that the majority seems to be able to say with any precision is that the regulation will likely contain some content-based exceptions. But as the San Francisco ordinance demonstrates, those exceptions are exceedingly rare. And when they do appear, they almost always involve content the State has determined is not suitable for children.

I fear that the majority’s rule will produce a generation of Americans—white, middle-class, suburban—who have no interest in, and no capacity to, engage in the kinds of play the First Amendment protects. And I fear that this Court, which prides itself on its willingness to protect the most vulnerable members of our society from harm, will be complicit in that generation’s continued marginalization. I respectfully dissent.’,

Decision: ‘deny’