CAMBRIDGE
UNIVERSITY PRESS

Beware the Lure of Narratives: "Hungry Judges" Should Not Motivate the Use of "Artificial Intelligence" in Law

Konstantin Chatziathanasiou[1]

[1]Institute for International and Comparative Public Law, University of Münster, Münster, Germany
Corresponding author: kchatzia@uni-muenster.de

Abstract

The "hungry judge" effect, as presented by a famous study, is a common point of reference to underline human bias in judicial decision-making. This is particularly pronounced in the literature on "artificial intelligence" (AI) in law. Here, the effect is invoked to counter concerns about bias in automated decision-aids and to motivate their use. However, the validity of the "hungry judge" effect is doubtful. In our context, this is problematic for, at least, two reasons. First, shaky evidence leads to a misconstruction of the problem that may warrant an AI intervention. Second, painting the justice system worse than it actually is becomes a dangerous argumentative strategy, as it undermines institutional trust. Against this background, this article revisits the original "hungry judge" study and argues that it cannot be relied on as an argument in the AI discourse or beyond. The case of "hungry judges" demonstrates the lure of narratives, the dangers of "problem gerrymandering," and, ultimately, the need for a careful reception of social science.

A. Introduction

We knew it all along, didn't we? A hungry judge is a stricter one. In 2011, a now famous study claimed to prove the point.[1] Judges were observed to issue harsher decisions just before their lunch break. The study demonstrated what legal realists had always been warning us about: judicial decision-making is not exempt from the pitfalls of human thinking.[2] Today the "hungry judge" effect has entered common wisdom.[3] Many commentators rely on it to argue for far-reaching consequences. Since the rise of technologies that are commonly summed up under the label

---

[1]Shai Danziger, Jonathan Levav & Liora Avnaim-Pesso, *Extraneous Factors in Judicial Decisions*, 108 PROC. NAT'L ACAD. SCIS. 6889 (2011).

[2]For a classic reference, see Joseph W. Bingham, *What Is The Law?*, 11 MICH. L. REV. 1, 21 (1912) ("To expect that judicial generalization and expression will not display these defects which are common in reasoning and language is to assume arbitrarily that our judges are all masters of human thought or to believe superstitiously that there is something in the judicial office which purifies and perfects the judge's reasoning."). For an example of a psychological perspective, see Chris Guthrie, Jeffrey J. Rachlinski & Andrew J. Wistrich, *Inside the Judicial Mind*, 86 CORNELL L. REV. 777 (2001). For a recent review article, see Gary Edmond & Kristy A. Martire, *Just Cognition: Scientific Research on Bias and Some Implications for Legal Procedure and Decision-Making*, 82 MOD. L. REV. 633 (2019) (arguing against "judicial exceptionalism").

[3]For references, see *infra* B.

"artificial intelligence" (AI),[4] machine-based decision aids appear as obvious remedies to human bias. After all, machines do not feel hunger or fatigue. The evidence for the fallibility of the "hungry judge," however, is not nearly as conclusive as commonly assumed. On the contrary, there are indications that the correlation found in the original study was spurious. Thus, reliance on the study to motivate legal interventions is problematic. First, it leads to a misconstruction of the normative problem. As automated decision-aids potentially introduce new biases to decision-making procedures, their employment requires a careful appraisal of risks and benefits. The "hungry judge" argument skews this analysis. Second, presenting the justice system as worse than it really is makes for a dangerous argumentative strategy. Trust in legal institutions is a precious societal resource, which should not be undermined without reliable evidence.[5]

This article consists of three parts. The first part introduces the "hungry judge" effect and describes its place in the legal literature, where it is broadly and uncritically used as an argument in the debate on the benefits and risks associated with the use of AI.[6] The second part questions this uncritical reliance.[7] In particular, it takes a close look at the "hungry judge" study and the arguments against its validity.[8] The last part reflects on our use of social science evidence in legal discourse.[9] It argues for a careful approach and warns against the lure of narratives and the dangers of "problem gerrymandering."

B. The "Hungry Judge" Effect and the Argument for Artificial Intelligence

The "hungry judge" effect is famous. It is the main result of an observational study by researchers Shai Danziger, Jonathan Levav, and Liora Avnaim-Pesso on the decisions of a real-life panel of Israeli judges ruling on the early release of prisoners on bail.[10] According to the study, the likelihood that an application for early release will be rejected increased as the breakfast or lunch break approached. The explanation is intuitive: fatigue or hunger influences our decision-making. The study was published prominently in the "Proceedings of the National Academy of Sciences" (PNAS) in 2011. In the same year, famous psychologist and Nobel laureate in economics, Daniel Kahneman, presented the study in his bestseller, "Thinking, Fast and Slow," thus making it accessible to a wider audience.[11] The process was apparently eased by the fact that Kahneman had already served as the study's "prearranged editor" at PNAS.[12] Due to the spectacular and intuitive result, the study received significant media coverage.[13] Since then, the study has been widely referenced, with Google Scholar counting well over one thousand citations.[14]

The study is cited in various contexts. In the legal literature, the study is typically used to confirm the basic assumption of legal realists—that factors beyond black-letter law influence a judge's decision and to motivate further research into these factors, with a special focus on heuristics and

---

[4]For a definition, see the HIGH-LEVEL EXPERT GROUP ON ARTIFICIAL INTELLIGENCE, ETHICS GUIDELINES FOR TRUSTWORTHY AI 36 (2019), https://www.aepd.es/sites/default/files/2019-12/ai-ethics-guidelines.pdf.

[5]*See also* Holger Spamann, Comment, *No, Judges Are Not Influenced by Outdoor Temperature (or Other Weather)*, (Harv. L. Sch., Discussion Paper No. 1036, 2020).

[6]*See infra* B.

[7]*See infra* C.

[8]For a previous short review in German, see Konstantin Chatziathanasiou, *Der hungrige, ein härterer Richter? Zur heiklen Rezeption einer vielzitierten Studie*, 74 JURISTENZEITUNG 455 (2019).

[9]*See infra* D.

[10]*See* Danziger et al., *supra* note 1.

[11]DANIEL KAHNEMAN, THINKING, FAST AND SLOW 43 (Farrar, Straus & Giroux eds., 2011). Kahneman was awarded the "Alfred Nobel Memorial Prize for Economics" in 2002 together with Vernon L. Smith, the pioneer of experimental economic research.

[12]*See* Danziger et al., *supra* note 1, at 6889. This procedure is now abolished. *See* Inder M. Verma, *Simplifying the Direct Submission Process*, 111 PROC. NAT'L ACAD. SCIS. 14311 (2014) (explaining that the process was intended for rare occasions and papers that required special attention, but that the majority of submissions in this process did not).

[13]Media references are abundant. For a prominent example, see John Tierney, *Do You Suffer From Decision Fatigue?*, N.Y. TIMES, (Aug. 17, 2011), https://www.nytimes.com/2011/08/21/magazine/do-you-suffer-from-decision-fatigue.html.

[14]Precisely, on Oct. 27, 2021, Google Scholar counted 1415 citations.

biases.[15] Thus, the study is used to emphasize that our knowledge on judicial decision-making is limited and that we should strive to learn more about it. However, the "hungry judge" effect is also used to argue for policy-interventions. Specifically, the "hungry judge" effect is often evoked in the discourse on automated decision-aids, or AI, in law. As we will see, reliance on the "hungry judge" is problematic in general, but particularly so in the latter context.

The use of AI in law is a sensitive issue. Generally, the label AI is used as an umbrella term for technologies that, based on the acquisition and interpretation of data, can perform, or "mimic," human decision-making tasks.[16] Such technologies have the potential to support legal practice. For example, AI can be employed for mass decision-making in administrative or tax proceedings.[17] The uses of AI, however, must be critically assessed from a human rights perspective.[18] Two prominent areas of application are particularly sensitive and, thus, are met with skepticism. The first area is algorithmic prediction of possible crime that allows for so-called predictive policing in order to guide the allocation of police resources.[19] The second area relates to risk-calculating tools that can be used to predict recidivism of defendants in criminal procedures.[20] The skepticism towards these applications is not just due to a general skepticism towards machine-based decisions.[21] The main worry about their use is that they may introduce new biases and risks to the decision-making procedure, particularly as they may disadvantage certain groups.[22]

Much of the critical debate focuses on the use of software such as COMPAS, a software that provides judges with a risk-assessment for an individual defendant in court. Such software is regarded as problematic for at least three reasons. First, the assessment provided by the software's algorithm may be biased. While it is possible to exclude obviously discriminating criteria—such as race—from the risk assessment, these criteria might still be correlated with other criteria that are

---

[15]*See, e.g.*, Daniel Bodansky, *Legal Realism and its Discontents*, 28 LEIDEN J. INT'L L. 267, 273 (2015); Julika Rosenstock, Tobias Singelnstein, & Christian Boulanger, *Versuch über das Sein und Sollen der Rechtsforschung*, in INTERDISZIPLINÄRE RECHTSFORSCHUNG: EINE EINFÜHRUNG IN DIE GEISTES- UND SOZIALWISSENSCHAFTLICHE BEFASSUNG MIT DEM RECHT UND SEINER PRAXIS 13 (Christian Boulanger, Julika Rosenstock & Tobias Singelnstein eds., 2019); Theodore Wilson, *The Promise of Behavioral Economics for Understanding Decision-making in the Court*, 18 CRIMINOLOGY & PUB. POL'Y 785, 797 (2019); Anna Sagana & D.A.G. van Toor, *The Judge as a Procedural Decision-Maker*, 228 ZEITSCHRIFT FÜR PSYCHOLOGIE 226, 228 (2020); Owen D. Jones, René Marois, Martha Farah & Henry Greely, *Law and Neuroscience*, 33 J. NEUROSCIENCE 17624, 17626 (2013); Anne Groggel, *The Role of Place and Sociodemographic Characteristics on the Issuance of Temporary Civil Protection Orders*, 55 L. & SOC'Y REV. 38, 42 (2021).

[16]*Cf.* HIGH-LEVEL EXPERT GROUP ON ARTIFICIAL INTELLIGENCE, ETHICS GUIDELINES FOR TRUSTWORTHY AI 36 (2019), https://www.aepd.es/sites/default/files/2019-12/ai-ethics-guidelines.pdf.

[17]*See, e.g.*, Yoan Hermstrüwer, *Artificial Intelligence and Administrative Decisions Under Uncertainty*, in REGULATING ARTIFICIAL INTELLIGENCE 199 (Thomas Wischmeyer & Timo Rademacher eds., 2020); Nadja Braun Binder, *Artificial Intelligence and Taxation: Risk Management in Fully Automated Taxation Procedures*, in REGULATING ARTIFICIAL INTELLIGENCE 295 (Thomas Wischmeyer & Timo Rademacher eds., 2020).

[18]*See* Lorna McGregor, Daragh Murray & Vivian Ng, *International Human Rights Law as A Framework for Algorithmic Accountability*, 68 INT'T & COMP. L. Q. 309 (2019); Alexander Tischbirek, *Artificial Intelligence and Discrimination: Discriminating Against Discriminatory Systems*, in REGULATING ARTIFICIAL INTELLIGENCE 103 (Thomas Wischmeyer & Timo Rademacher eds., 2020); Mireille Hildebrandt, *Law as Information in the Era of Data-Driven Agency*, 79 MOD. L. REV. 1 (2016) (arguing that law scholars should engage in the new grammar of data driven law in order to safeguard human freedom and autonomy).

[19]LUCIA M. SOMMERER, PERSONENBEZOGENES PREDICTIVE POLICING. KRIMINALWISSENSCHAFTLICHE UNTERSUCHUNG ÜBER DIE AUTOMATISIERUNG DER KRIMINALPROGNOSE (Nomos, 2020); Timo Rademacher, *Artificial Intelligence and Law Enforcement*, in REGULATING ARTIFICIAL INTELLIGENCE 225 (Thomas Wischmeyer & Timo Rademacher eds., 2020).

[20]On 'evidence-based sentencing' and actuarial risk assessments, see Sonja B. Starr, *Evidence-based Sentencing and the Scientific Rationalization of Discrimination*, 66 STAN. L. REV. 803 (2014) (with particular references to the scores generated by the software products LSI-R and COMPAS).

[21]On possible bias *against* machines, see CÉSAR A. HIDALGO, DIANA ORGHIAN, JORDI ALBO CANALS, FILIPA DE ALMEIDA & NATALIA MARTIN, HOW HUMANS JUDGE MACHINES (2021) (showing that judging machines is outcome-based, while judging humans accounts for human intentions).

[22]*See, e.g.*, Aziz Z. Huq, *Racial Equity in Algorithmic Criminal Justice*, 68 DUKE L.J. 1043 (2019).

very well included.[23] Second, it is difficult to control the inner workings of such software, as they are commercial and proprietary products.[24] Third, there are dangers connected to the interaction between the automated decision-aid and human judges who might place too much confidence in the recommendation they are provided with.[25] Warnings that are meant to counteract overreliance have been found to have only a very limited effect.[26]

These problems are well-known and have a prominent place in the discourse on AI in law. This is where the "hungry judge" effect comes into play as a counterargument. In particular, the "hungry judge" effect is used to underline that human judgment is *already* susceptible to biases.[27] To give an example, the U.S.-based economist Jennifer Doleac who works on issues of sentencing and on racial discrimination, counters the discussion about biased algorithms as follows: "[T]hat discussion misses an important point: Humans are biased. We routinely get things wrong due to an array of cognitive biases. Judges' decisions are affected by factors such as whether they're hungry (decisions are far less favorable just before lunch) and if their local football team won that weekend[.]"[28] So, the argument goes, even if there is a danger that algorithmic decision-making was biased, *could it be* more biased than judges who leave people incarcerated because they cannot control their hunger? AI systems could, at least, help judges detect, counteract, or exclude their own subconscious biases, thus increasing fairness and legal certainty.[29]

Such reference to "hungry judges" is not just occasional but can be seen as particularly popular—and indeed common. References can be found in all different genres of writing on AI in law: In the scholarly literature,[30] in policy papers,[31] as well as in opinion pieces.[32] The references are so broad that it feels safe to describe "hungry judges" as part of an "AI lore." But as

[23]*See* Starr, *supra* note 20, at 812 (criticizing the incorporation of socioeconomic status and demographic categories, ranging from high school grades to neighborhood crime rate, within the relevant variables).

[24]*Id.*

[25]*See, e.g.*, John Zerilli, Alistair Knott, James Maclaurin & Colin Gavaghan, *Algorithmic Decision-Making and the Control Problem*, 29 MINDS & MACHINES 555 (2019) (warning about "'the control problem' understood as the tendency of the human within a human-machine control loop to become complacent, over-reliant or unduly diffident when faced with the outputs of a reliable autonomous system").

[26]Christoph Engel & Nina Grgić-Hlača, *Machine Advice with a Warning About Machine Limitations: Experimentally Testing the Solution Mandated by the Wisconsin Supreme Court*, 13 J. LEG. ANAL. 284 (2021).

[27]But then again, on the risk-magnifying potential of AI, see Thomas G. Dietterich, *Robust Artificial Intelligence and Robust Human Organizations*, 13 FRONT. COMPUT. SCI. 1 (2019).

[28]Jennifer Doleac, *Let Computers Be the Judge: The Case for Incorporating Machine Learning into the U.S. Criminal Justice Process*, MEDIUM (Apr. 20, 2017), https://medium.com/@jenniferdoleac/let-computers-be-the-judge-b9730f94f8c8.

[29]*See, e.g.*, Jasper Ulenaers, *The Impact of Artificial Intelligence on the Right to a Fair Trial: Towards a Robot Judge?*, 11 ASIAN J. L. ECON. 1, 24–25 (2020); Songül Tolan, *Fair and Unbiased Algorithmic Decision Making: Current State and Future Challenges* 6 (JRC Technical Reports, Working Paper No. 2018-10, 2018).

[30]*See, e.g.*, Tania Sourdin, *Judge v Robot? Artificial Intelligence and Judicial Decision-Making*, 41 UNIV. N.S.W. L. J. 1114, 1128 (2018); Vincent Chiao, *Fairness, Accountability and Transparency: Notes on Algorithmic Decision-making in Criminal Justice*, 15 INT'L. J. L. CONTEXT 126, 132 (2019); Kenneth Tung, *AI, the Internet of Legal Things, and Lawyers*, 6 J. MGMT. ANALYTICS 390, 395 (2019); Aleš Završnik, *Algorithmic Justice: Algorithms and Big Data in Criminal Justice Settings*, EUR. J. CRIMINOLOGY 1, 11 (2019); Anna C. F. Lewis, *Where Bioethics Meets Machine Ethics*, 20 AM. J. BIOETHICS 22, 23 (2020); Thomas Grote & Ezio Di Nucci, *Algorithmic Decision-Making and the Problem of Control*, *in* TECHNOLOGY, ANTHROPOLOGY, AND DIMENSIONS OF RESPONSIBILITY 89 (Birgit Beck & Michael Kühler eds., 2020); SOMMERER, *supra* note 19, at 171; Mirko Bagaric, Dan Hunter & Nigel Stobbs, *Erasing the Bias Against Using Artificial Intelligence to Predict Future Criminality: Algorithms Are Color Blind and Never Tire*, 88 UNIV. CIN. L. REV. 1037, 1077 (2019).

[31]*See, e.g.*, KONRAD LISCHKA & ANITA KLINGEL, WENN MASCHINEN MENSCHEN BEWERTEN: INTERNATIONALE FALLBEISPIELE FÜR PROZESSE ALGORITHMISCHER ENTSCHEIDUNGSFINDUNG 11 (2017), https://www.bertelsmann-stiftung.de/fileadmin/files/BSt/Publikationen/GrauePublikationen/ADM_Fallstudien.pdf; FILIPPO RASO, HANNAH HILLIGOSS, VIVEK KRISHNAMURTHY, CHRISTOPHER BAVITZ & LEVIN KIM, ARTIFICIAL INTELLIGENCE & HUMAN RIGHTS: OPPORTUNITIES & RISKS 20 (2018); COLIN GAVAGHAN, ALISTAIR KNOTT, JAMES MACLAURIN, JOHN ZERILLI & JOY LIDDICOAT, GOVERNMENT USE OF ARTIFICIAL INTELLIGENCE IN NEW ZEALAND 39 (2019), https://www.otago.ac.nz/caipp/otago711816.pdf. *See also* Tolan, *supra* note 29, at 6.

[32]*See, e.g.*, Bryce Goodman & Seth Flaxman, *European Union Regulations on Algorithmic Decision Making and a "Right to Explanation,"* 38 AI MAG. 50, 46; Doleac, *supra* note 28.

already indicated, referring to "hungry judges" is not just an innocent triviality. The "hungry judge" argument matters. The introduction of AI into legal decision-making comes with potential trade-offs. This requires political debate and a careful cost-benefit analysis. The "hungry judge" argument paints a very negative picture of the current reality in the courtroom. Compared to such arbitrary practice as letting a judicial decision depend on its proximity to a break, every intervention that makes a decision more rational appears legitimate—even urgent. Thus, the argument increases the appeal of automated decision-aids to an audience that might be critical at first.

The broad reliance on the "hungry judge" effect and the sensitive context of far-reaching policy proposals warrants a critical engagement with the evidence. As we will see, the confidence that the discourse puts into the original study is not well placed.[33] So, while human bias might indeed be a valid argument for the use of automated decision-aids, the particular reference to the "hungry judge" effect is not.

C. Is There a "Hungry Judge" Effect?

The following discussion starts by revisiting the original study in detail[34] and then moves to the arguments against its validity. The objections focus on the magnitude of the found effect,[35] the factual basis of the study,[36] and alternative explanations for the found effect.[37]

I. Danziger et al.: "Yes"

In their 2011 study, Danziger et al. analyze 1,112 decisions by eight judges on fifty days over the course of ten months.[38] The judges decided on requests from prison inmates to have their prison sentences suspended. During the examined period, between fourteen and thirty-five cases were processed every day, with an average processing time of just under six minutes per case. The authors' data include the order of the cases and the time of the hearing. Each day, sessions were structured around two breaks: A (late) breakfast and a lunch break. According to the authors, the judges had no knowledge of the next case and no influence on the order of the cases. On this basis, Danziger et al. test how the position of a case in the order of all cases and breaks affect the decision-making behavior of the judges. All case decisions were coded as either favorable or unfavorable. The hypothesis of the study is that, due to fatigue, repeated decisions would lead to more decisions maintaining the status quo, especially because favorable decisions were more labor-intensive due to the greater effort involved in giving a justification for an early release.[39]

The analysis shows that the number of unfavorable decisions strongly increased before the breaks. The highest probability of a favorable decision is around sixty-five percent at the beginning of a session and then falls to almost zero—after a break, it is then again around sixty-five percent.[40]

---

[33]For authors who do reference the critiques, see Marion Oswald, *Algorithm-assisted Decision-making in the Public Sector: Framing the Issues Using Administrative Law Rules Governing Discretionary Power*, 376 PHILO. TRANS. R. SOC. A.: MATH., PHYS. & ENGIN. SCI. 6 (2018), https://royalsocietypublishing.org/doi/full/10.1098/rsta.2017.0359; Mario Martini, *Regulating Algorithms: How to Demystify the Alchemy of Code?*, in ALGORITHMS & L. 105 (Martin Ebers & Susana Navas eds., 2020).

[34]See *infra* I.

[35]See *infra* II.

[36]See *infra* III.

[37]*See infra* IV.

[38]Danziger et al., *supra* note 1, at 6889.

[39]Danziger et al., *supra* note 1, at 6890 ("Two indicators support our view that rejecting requests is an easier decision—and, thus, a more likely outcome—when judges are mentally depleted: (i) favorable rulings took significantly longer (M = 7.37 min, SD = 5.11) than unfavorable rulings (M = 5.21, SD = 4.97), t = 6.86, P < 0.01, and (ii) written verdicts of favorable rulings were significantly longer (M = 89.61 words, SD = 65.46) than written verdicts of unfavorable rulings (M = 47.36 words, SD = 43.99), t = 12.82, P < 0.01.").

[40]*Id.*

Using a statistical control technique,[41] the authors examine to what extent other factors—such as certain characteristics of the case—could explain these results. Of the "legal" variables that can be found in the case file, previous convictions and the lack of a rehabilitation program significantly reduce the likelihood of a favorable decision. Importantly, the order in which cases were decided was found to be decisive.

Crucially, the internal validity of the study—that is, whether the research design justifies the inference—depends on whether the sequence of cases was actually random—that is, whether it can be safely assumed that cases that appeared at the end of a session did not differ in any particular characteristic from the earlier cases.[42] Only then does the conclusion hold that it is, indeed, the order that explains a certain decision. Instead, this would not be the case if an unobserved variable affected the order of the cases. In contrast to an experiment, in an observational study, the order in which cases are decided and the information available to the decision-makers cannot be determined by the experimenter, thus the influence of one or more unobserved variables cannot be excluded.[43]

Every observational study that aims at establishing causality tries to exclude—or control for—unobserved influencing factors. Here, the assumption that the judges had no knowledge of the next case is of crucial significance. The same applies to the statement that the time at which each case is dealt with depends only on the arrival time of an inmate's lawyer.[44] Based on these assumptions, the order of the cases would be exogenous from the judges' perspective. Thus, Danziger et al. conclude that the only explanation for the high rejection rate shortly before the breaks is the judges' fatigue—or, more bluntly, their hunger.

II. Effect Size and Implausibility

There are several objections to the validity of the "hungry judge" study. The most general one is related to the size of the effect found by Danziger et al. The effect is particularly large, which calls for caution.[45] Generally, psychological effects reported for the first time tend to be larger than in later replications.[46] But the effect size also appears very large specifically when compared to the effects measured in controlled laboratory studies on fatigue symptoms.[47] For this reason, critics like psychologist Daniel Lakens have even dismissed the study outright. On his blog, Lakens describes the effect as implausible and even impossible. It is simply too big to be caused by a psychological mechanism.[48] To quote Lakens: "If hunger had an effect on our mental

[41] A logistic regression with a fixed effects model for the individual judges in order to take their specific tendencies into account.

[42] Such a random order would approximate the experimental ideal of random assignment, that is, the random distribution of the experimental treatment between an experimental group and a control group. For an introduction to experimental methods in law, see Colin Camerer & Eric Talley, *Experimental Study of Law*, *in* HANDBOOK OF LAW AND ECONOMICS, II (Polinsky & Shavell eds., 2007); Konstantin Chatziathanasiou & Monika Leszczyńska, *Experimentelle Ökonomik im Recht*, 8 RECHTSWISSENSCHAFT 314 (2017).

[43] But see the experiment by Holger Spamann & Lars Klöhn, *Justice Is Less Blind, and Less Legalistic, Than We Thought: Evidence From an Experiment with Real Judges*, 45 J. LEG. STUD. 255 (2016), in which the authors themselves determine which information is available to the judges in a fictitious case that is modeled after a real one.

[44] Danziger et al., *supra* note 1, at 6892.

[45] *See* Andreas Glöckner, *The Irrational Hungry Judge Effect Revisited: Simulations Reveal That the Magnitude of the Effect is Overestimated*, 11 JUDGMENT AND DECISION MAKING 601, 602 (2016); Daniel Lakens, *Impossibly Hungry Judges,* THE 20% STATISTICIAN (July 3, 2017), http://daniellakens.blogspot.com/2017/07/impossibly-hungry-judges.html.

[46] *Id.* at 608. *See also* Open Science Collaboration, *Estimating the Reproducibility of Psychological Science*, 349 SCI. 943 (2015).

[47] Glöckner, *supra* note 45, at 602 ("It might, however, be argued that manipulations of depletion and exhaustion might be stronger in reality than in the lab causing stronger effects.").

[48] *See* Lakens, *supra* note 45 (emphasizing that the effect size in the study has a Cohen's *d* of nearly two—for comparison, this amounts to the effect size when comparing the "difference between the height of 21-year old men and women in The Netherlands").

resources of this magnitude, our society would fall into minor chaos every day at 11:45. Or at the very least, our society would have organized itself around this incredibly strong effect of mental depletion."[49] The effect appears so big that one would actually not need a scientific study to identify it. For Lakens, the lack of a plausible theory that explains an effect of this magnitude already suffices to dismiss the study entirely. Another good reason for this objection is that we do not observe any kink in the data immediately after a lunch break, where our attention typically fades and we would assume another point of possible mental depletion.[50] Apart from this "implausibility critique," there are further objections. The next one attacks the factual basis of the study.

III. Factual Basis: Non-random Ordering and Legal Representation

Danziger et al. rely crucially on the assumption that the order of the cases is random and, thus, exogenous to the decision-making process. This assumption has been forcefully challenged. For a short and very critical reply in PNAS, Keren Weinshall-Margel and John Shapard analyzed the data of the original study—as well as other self-collected data—and conducted additional interviews with the court personnel involved.[51] They point out that the order of the cases is *not* random: The panel tries to deal with all cases from one prison before a break, before then moving to the cases of the next prison after a break. Most importantly, though, requests from prisoners who are not represented by a lawyer are typically dealt with at the end of each session. So, prisoners without legal representation are less likely to receive a favorable decision compared to those with legal representation.[52] Additionally, lawyers often represent several inmates and decide on the order in which the cases are presented—it might well be possible that they start with the strongest cases.[53]

Danziger et al. reacted to this criticism in a short reply, seeing no reason to revise their conclusions. But, the reply does not appear conclusive. To begin with, they state that ordering did not depend on the prison of origin. However, to back that claim, they could only provide data on the prison of origin for five of the fifty days they used for their first analysis. Further, they rule out an ordering according to specific prison because the order of the cases depended on the arrival of the lawyers.[54] In addition, the cases in which the same lawyer represented several inmates were too rare to affect the analysis. Most importantly though, Danziger et al. report to have now included legal representation as a control variable in their regression analysis, and that they see their results confirmed.[55] However, they do not report whether this reduces the size of the effect. Thus, their answer is vague. Crucial questions remain unanswered: how strong is the effect of ordering if one accounts for legal representation? And, what is even more important: Can it be said with certainty that the results are not influenced by an unobserved variable?

---

[49]*Id.*

[50]*Id.*

[51]Keren Weinshall-Margel & John Shapard, *Overlooked Factors in the Analysis of Parole Decisions*, 108 Proc. Nat'l Acad. Scis. E833 (2011).

[52]*Id.* ("Using the same decision rules as Danziger et al., our data indicate that unrepresented prisoners account for about one-third of all cases, but they prevail only 15% of the time, whereas prisoners with counsel prevail at a 35% rate.").

[53]*Id.* ("We suspect that attorneys present their best cases first and save their weakest cases for last, adding to the downward trend of prisoner success.").

[54]Shai Danziger, Jonathan Levav, & Liora Avnaim-Pesso, *Reply to Weinshall-Margel and Shapard: Extraneous Factors in Judicial Decisions Persist*, 108 Proc. Nat'l Acad. Scis. E834 (2011).

[55]*Id.* ("The original results replicate in every analysis; case order and the meal break remain robust predictors of the parole decision. The presence of legal counsel was always positively correlated with release likelihood but was not a significant predictor in every single model that we ran.").

IV. Alternative Explanation: Time Management

As the size of the effect appears implausible, and the factual basis for assuming a random case order unsure, could there be alternative explanations for the observed effect? Apart from the already mentioned legal representation, a further unobserved variable that could influence the case order and explain the results would be the workload management by the judges.[56] Indeed, there are some clues that time management could play a role. According to Danziger et al., rendering favorable decisions takes more time than rendering unfavorable ones. Thus, judges might want to avoid starting more complicated cases before the break, as they would run the risk of exceeding their time limit and take their break later or not at all.[57] To assess whether the next case is a time-consuming one, superficial hints can be enough—just think of the thickness of the file, a post-it from a clerk, or the wording of the applicant's request. Such a planning decision could lead to a decreased likelihood that positive cases are among the cases with a higher—or later—order within a session. Could this be a possible explanation for the effect that we see before the breaks?

Social psychologist Andreas Glöckner came up with an elegant way to test this alternative explanation. Glöckner uses a computer simulation. This method is complementary to an observational approach: Instead of observing and analyzing existing data, a simulation generates data. Thus, it answers the question what data would look like under certain specified conditions. The results can then be contrasted. In his argument, Glöckner makes the assumption that the judges, or their staff, have external clues as to whether a case will take more or less time. This assumption is plausible and has grave effects.

To get an idea of how strong the effect of a planning decision is, Glöckner simulates the decisions of a judge who is not subject to errors and distortions due to hunger or fatigue. How would such a judge handle her workload under time-constraints? Using a statistical program,[58] Glöckner generates 10,000 randomly ordered cases, of which thirty-six percent are positive and sixty-four percent negative, thus mirroring the proportions in the original study. Then, he generates response times that follow the averages reported by Danziger et al. These averages vary for positive and negative rulings with the positive rulings taking significantly more time.[59] To mirror that a judge does not have unlimited time at her disposal, Glöckner sets a time limit for decisions at sixty minutes. In the simulation, the judge handles cases until the next case would go beyond the time limit. This models the situation in which a judge would recognize from an external clue that the next case cannot be dealt with without going over a time limit. The data that the simulation generates can then be contrasted with the observations from the original study.

It turns out that *even* in the case of a judge to whom cases are presented in random order, the probability of positive decisions decreases sharply towards the end of the session. The closer to the end of the session, the fewer positive cases appear.[60] The explanation is straightforward: Because negative cases can be dealt with in a shorter timeframe, the proportion of cases that can still be dealt with in the remaining time before the break is greater among the negative cases. This reduces the likelihood that a session will end with a favorable decision—leading to a bias in the sample. At

---

[56]From an econometric perspective, the problem can also be described as simultaneity bias in the regression model. An independent variable (treatment variable) might be correlated with the confounding variable (error term). In our case, there could be a correlation between the order of the cases or the timing of the break and the selection of the cases observed in the study. The correlation could be established through time management, which acts as an unobserved variable on both.

[57]Just imagine this dreadful scenario.

[58]*See* Glöckner, *supra* note 45, at Appendix (providing an exemplary STATA code).

[59]*See* Danziger et al., *supra* note 1, at 6890 ("Two indicators support our view that rejecting requests is an easier decision—and, thus, a more likely outcome—when judges are mentally depleted: (i) favorable rulings took significantly longer (M = 7.37 min, SD = 5.11) than unfavorable rulings (M = 5.21, SD = 4.97), t = 6.86, P < 0.01, and (ii) written verdicts of favorable rulings were significantly longer (M = 89.61 words, SD = 65.46) than written verdicts of unfavorable rulings (M = 47.36 words, SD = 43.99), t = 12.82, P < 0.01.").

[60]Glöckner, *supra* note 45, at 604.

the same time, this also explains why the probability of a favorable decision is higher at the beginning of a session right *after* a break. It is more likely that a session will end on an unfavorable decision, with a favorable decision being postponed to the beginning of the next session.[61] Thus, a real judge would generate a similar pattern, if only she operates under a time limit and has a rough clue about whether she will have to write a short or long opinion. This holds even if her judgment is not distorted by hunger or fatigue.

Yet, Glöckner identifies another problem. Since the sessions in the original study were of different lengths, Danziger et al. compared only the first ninety-five percent of the cases. Could this "censoring" of data also have biased the sample? Glöckner simulates a judge who has no information about the complexity of the case but has to stop when a time limit is reached. As positive cases take longer to process, the likelihood increases that it is a positive case on which time will run out. Thus, the decision by Danziger et al. to disregard the cases on which the sessions end, affects positive cases more than negative ones. This increases the relative frequency of negative cases in the sample. This censoring of the last 5% of the observations also intensifies the downward trend of favorable decisions in the previous scenarios of a judge who makes planning decisions.[62]

The simulations show that an explanation other than hunger or fatigue is indeed likely. If we accept the plausible assumption that at least some planning occurs, we can explain the results also on the basis of a judge who is not subject to any distortion.[63] This could then also be described as a case of reverse causality. It is not the position of the case in the order of all cases that turns the case into a positive or negative one. It is much rather its characteristic as positive or negative that makes a case more likely to appear at a certain position in the order of all cases. This explanation is not sensational but very plausible. So, given the doubts about the magnitude of the effect, the factual basis on the assumption of a random order, and the plausibility of time management as an alternative explanation, it is safe to say that the results of the study are not credible. In sum, based on the discussed evidence, there is no "hungry judge" effect.

D. The Delicate Reception of Social Science Evidence in Law

Notwithstanding the strong doubts about the study's validity, the "hungry judge" study is an uncomfortable case for legal scholarship for several reasons. First and foremost, it is a study that —from the outside—appears to be carried out thoughtfully and rigorously. Then, while the effect size is implausible, effect sizes are not something legal scholars are trained to discuss. Importantly, most legal scholars must rely on other secondary clues. The study was published in a highly reputed outlet and endorsed by the most famous authority on cognitive bias research.[64] Thus, it carries the secondary signals that legal scholars typically—and with good reason—rely upon when appraising the output of a research field that is not theirs.[65] Ultimately, what makes the study especially hard to deal with is that it confirms an intuition, or popular narrative, which is already shared. Possibly, our wish to see our intuition confirmed is even stronger than our lawyerly caution. I see three main takeaways from this case that relate, first, to research on judicial bias;[66] second, to the *appraisal* of social science from a legal perspective[67]; and, third, to the problem of *selective* presentation and reception of social science evidence in the AI discourse and beyond.[68]

---

[61]Glöckner, *supra* note 45, at 605, 608 (conceding that this cannot explain the high values at the beginning of the day).
[62]Glöckner, *supra* note 45, at 606.
[63]Glöckner, *supra* note 45, at 607.
[64]KAHNEMAN, *supra* note 11, at 43.
[65]On the associated dangers, see Niels Petersen & Konstantin Chatziathanasiou, *On the Seductions of Quantification: A Rejoinder,* 19 INT'L J. CONST. L. 1854 (2021).
[66]See *infra* I.
[67]*See infra* II.
[68]*See infra* III.

I. Research on Judicial Bias

The first main takeaway relates to the research on influence of non-legal factors on judicial deci-sion-making. The state of this research should be treated with caution. On the one hand, there is a solid body of research on the influence of certain factors like, for example, ideology or demo-graphics on the decision-making of apex courts, with a particular focus on the Supreme Court of the United States.[69] This body of research can be described as well-developed and robust. The Supreme Court has, for a long time, been at the center of scholarly attention in political sci-ence scholarship, and many empirical studies provide evidence to support theories on judicial behavior.[70] On the other hand, there is research that looks into the influence of variables on judi-cial-decision making that is not as strongly grounded in theory, but rather mirror intuitions.[71] Furthermore, some of this research appears to primarily be after sensational results. Apart from the effect of hunger on parole decisions, this research examines the effects of sleepiness[72] or sport results[73] on sentencing, or weather on asylum decisions.[74] While this research is valuable, it cannot be considered as robust as research on ideology because it wrestles with methodological problems. Scholars like Holger Spamann have regularly dismantled such approaches.[75] Spamann himself has taken a different route in his research. In a study co-authored by Lars Klöhn, they examine the effects of legally relevant and irrelevant information about a defendant on the assessment of legal questions in an experiment with actual judges, thus having full control of the information that the judges receive.[76] This line of research, which has recently been substantially extended,[77] not only covers potential non-legal influences, but also the influence on intra-legal factors on the tech-niques of judicial decision-making. All this is certainly not meant to say that judicial decision-making is not susceptible to bias or error,[78] but sensational studies that sound too good to be true are better met with a healthy dose of skepticism. Thus, should such a sensational study matter for a

---

[69]*See* Jeffrey A. Segal & Albert D. Cover, *Ideological Values and the Votes of US Supreme Court Justices*, AM. POL. SCI. REV. 557 (1989); Jeffrey A. Segal, Charles M. Cameron & Albert D. Cover, *A Spatial Model of Roll Call Voting: Senators, Constituents, Presidents, and Interest Groups in Supreme Court Confirmations*, AM. J. POL. SCI. 96 (1992); Jeffrey A. Segal, Lee Epstein, Charles M. Cameron & Harold J. Spaeth, *Ideological Values and the Votes of US Supreme Court Justices Revisited*, 57 J. POLIT. 812 (1995); Christina L. Boyd, Lee Epstein & Andrew D. Martin, *Untangling the Causal Effects of Sex on Judging*, 54 AM. J. POL. SCI. 389 (2010).

[70]For an overview, see Niels Petersen & Konstantin Chatziathanasiou, *Empirische Verfassungsrechtswissenschaft. Zu Möglichkeiten und Grenzen quantitativer Verfassungsvergleichung und Richterforschung*, 144 ARCHIV DES ÖFFENTLICHEN RECHTS 501, 533 (2019).

[71]I thank an anonymous reviewer for raising this point.

[72]*See* Kyoungmin Cho, Christopher M. Barnes & Cristiano L. Guanara, *Sleepy Punishers Are Harsh Punishers: Daylight Saving Time and Legal Sentences*, 28 PSYCHOL. SCI. 242 (2017) (finding that longer prison sentences will be imposed immediately after the switch to summer time and the resulting reduced sleep). For a critical response see Holger Spamann, *Are Sleepy Punishers Really Harsh Punishers? Comment on Cho, Barnes, and Guanara (2017)*, 29 PSYCH. SCI. 1006, 1008 (2018) (showing that the effect disappears if measurement of judicial severity includes cases in which no prison sentence is pronounced).

[73]*See, e.g.*, Ozkan Eren & Naci Mocan, *Emotional Judges and Unlucky Juveniles*, 10 AM. ECON. J. APPL. ECON. 171 (2018) (suggesting that there are tougher juvenile sentences after a surprising defeat of the local football college team).

[74]*See, e.g.*, Anthony Heyes & Soodeh Saberian, *Temperature and Decisions: Evidence from 207,000 Court Cases*, 11 AM. ECON. J. APPL. ECON. 238 (2019) (finding that a ten degrees Fahrenheit increase in case-day temperature leads to a reduction of US immigration judges' decisions favorable to the applicant by 6.55 percent). For a critique, see Holger Spamann, *supra* note 5 (failing to replicate the result with an enlarged sample).

[75]*See* Spamann, *supra* note 72. *See also* Spamann, *supra* note 5.

[76]Spamann & Klöhn, *supra* note 43.

[77]Holger Spamann, Lars Klöhn, Christophe Jamin, Vikramaditya Khanna, John Zhuang Liu, Pavan Mamidi, Alexander Morell & Ivan Reidel, *Are There Common/Civil Law Differences and Precedent Effects in Judging Around the World? A Lab Experiment* (Harv. L. Sch., Discussion Paper No. 1044, 2020).

[78]*See* Alma Cohen & Crystal S. Yang, *Judicial Politics and Sentencing Decisions*, 11 AM. ECON. J. ECON. POL'Y 160 (2019) (finding that Republican-appointed judges issue longer sentences for black defendants than similar non-blacks).

policy argument, we would be well served to check how the study has been received within the relevant community. While much of this extra-legal disciplinary knowledge might be implicit, recent years have made debates in other disciplines much more accessible. Podcasts and blogs have played their part in this development, as the aforementioned example of Daniel Lakens' work illustrates.

II. Appraising Social Science From a Legal Perspective

Beyond the specific case of research on bias in judicial decision-making and with regard to social science studies more generally, a delicate balance must be struck. In particular, two extremes should be avoided: while a single study should not be overestimated as reliable evidence, neither should its useful potential to alert us be ignored. Still, the latter potential must be put into perspective.

The problems of the "hungry judge" effect already show why we would be ill-advised to justify a legal reform with a single study.[79] It cannot be ruled out that later studies refute primary findings or, at least, put their informative value into perspective. This leads to the next question: when does social science evidence suffice for justifying a legal intervention? This is a demanding normative question, especially as legal scholars are typically not trained in social science methods.[80] A plausible criterion or starting point would be whether we can regard scientific knowledge as consolidated. Such consolidation takes place generally through replication,[81] meta-analysis of several studies,[82] or even through the growth of a field that leads to a denser discursive control through peers. The question of if the consolidation meets the standards required for justifying a particular legal intervention remains a question legal scholars have to answer.[83]

However, one could say that studies like the one discussed, provide potential to challenge us. This potential might indeed be valuable. It makes us question whether we are as immune to bias or as objective as we would like—or as the law requires us—to be.[84] Indeed, the study by Danziger et al. might be used solely as a rhetorical device, without founding a policy-argument on it.[85] Still, we should acknowledge that even such rhetorical use is not harmless. Even if a certain irritating effect might be welcome, this does not justify spreading studies which are problematic from a methodological standpoint and which "paint the justice system worse than it actually is."[86] The public perception of the justice system matters. Exaggerations can undermine trust in the legal system. Such trust, however, is a precious societal resource, as the judiciary critically depends on public support to maintain its independence. As readers of this journal are aware, judicial

---

[79]For a warning, see Susan McNeeley & Jessica J. Warner, *Replication in Criminology: A Necessary Practice*, 12 EUR. J. CRIMINOLOGY 581, 591 (2015) (also referring to the "hungry judge" study). *See also* Richard Lempert, *Empirical Research for Public Policy: With Examples from Family Law*, 5 J. EMPIRICAL LEGAL STUD. 907, 925 (2008).

[80]On the danger that methodological caveats are lost in translation, see Niels Petersen & Konstantin Chatziathanasiou, *supra* note 65.

[81]For a path-breaking example, see Open Science Collaboration, *supra* note 46. From the perspective of criminology, see McNeeley & Warner, *supra* note 79. For recent replication studies in experimental law and economics, see Svenja Hippel & Sven Hoeppner, *Biased judgements of Fairness in Bargaining: A Replication in the Laboratory*, 58 INT'L REV. L. ECON. 63 (2019); Svenja Hippel & Sven Hoeppner, *Contracts as Reference Points: A Replication*, 65 INT'L REV. L. ECON. 105973 (2021).

[82]For a recent example, see Piotr Bystranowski, Bartosz Janik, Mciej Próchnicki & Paulina Skórska, *Anchoring Effect in Legal Decision-making: A Meta-analysis*, 45 L. HUM. BEHAV. (2021).

[83]*See, e.g.*, Christoph Engel, *Empirical Methods for the Law*, 174 J. INST. THEORETICAL ECON. 5, 10 (2018) (arguing for a closer match between legal research question and empirical method).

[84]*See* Edmond & Martire, *supra* note 2.

[85]*See, e.g.*, Jan Christoph Bublitz, *What is Wrong with Hungry Judges? A Case Study of Legal Implications of Cognitive Science*, in LAW, SCI. & RATIONALITY (Waltermann, Roef, Hage & Jelicic eds., 2020) (arguing that decisions as described in the "hungry judge" study "follow from features of the law such as indeterminacy, reasonable disagreement, and a diverse judiciary").

[86]*See also* Spamann, *supra* note 5, at 12.

independence is currently under pressure in many countries.[87] Studies like the one discussed are easily exploited politically. In fact, German far-right populists have already used the "hungry judge" effect to demand less funding—a "diet"—for judges and more severe sentences for criminals.[88]

III. The Problem of Selective Reception of Social Science Evidence

In court, the establishment of the facts is a highly contested matter. Each side presents their version of the relevant events to make a favorable judgment more likely. Similarly, in the context of legal or policy choices, the danger of a selective reception of social science evidence looms large.

The reasons for selective reception may be several. Regarding the "hungry judge" effect, our preference for narratives that confirm what we already thought is on full display. Psychologists might even speak of "confirmation bias" or even "motivated reasoning."[89] The problem may not appear serious at first sight, but we should be aware that our responsiveness to narratives may be exploited. Like in court, the underlying strategy is a selective presentation of the facts that are supposed to constitute a problem—"problem gerrymandering," so to speak.[90] A study like the one discussed can be employed to frame a problem in a certain way, thus making a particular, favored intervention more likely.[91] Critically, problem construction and proposed solutions are typically entangled.[92] As the "hungry judge" study has become part of the debate on "AI and law," there is a danger that we will come to wrong conclusions on the necessity of interventions, or that we will possibly choose the wrong ones. Imagine that—based on the "hungry judge" study—Israeli legislators had decided to entrust the decision of bail to an algorithm. This would have created new risks[93] without even fixing an existing problem.

E. Conclusion

The "hungry judge" effect is famous, but—as we have seen—there are strong doubts about its validity. Therefore, we should not rely on it to motivate the introduction of automated decision-aids nor other policy-interventions.[94] The case presented, here, should also make us vigilant regarding our use of social science evidence more generally. Especially when an empirical study fits our preferred narrative but at the same time sounds too good to be true, caution is advisable. Similarly, at the end of this article, we should not be too sweeping with our conclusions. Simply

---

[87]On past and present challenges, see Hans Petter Graver & Peter Čuroš, *Judges Under Stress: Understanding Continuity and Discontinuity of Judicial Institutions of the CEE Countries*, 22 GERMAN L.J. 1147 (2021).

[88]In a Facebook post, the far-right German party "Alternative für Deutschland" referenced the "hungry judge" study arguing to put judges on a diet as a remedy to "cuddle justice," Apr. 30, 2018, screenshot on file with author ("Mit hungrigen Richtern gegen die weichgespülte Kuscheljustiz").

[89]On the phenomenon, see Ziva Kunda, *The Case for Motivated Reasoning*, 108 PSYCH. BULL. 480 (1990) (proposing "that motivation may affect reasoning through reliance on a biased set of cognitive processes—that is, strategies for accessing, constructing, and evaluating beliefs").

[90]*Cf.* Steve Woolgar & Dorothy Pawluch, *Ontological Gerrymandering: The Anatomy of Social Problems Explanations*, 32 SOC. PROBS. 214 (1985).

[91]To be clear: I am not saying that all of the scholarly literature on AI is consciously pursuing a strategy of problem gerrymandering. Quite a few papers that mention the "hungry judge" effect also appreciate the risks of AI in law.

[92]On this a basic insight from science studies, see, e.g., JULIA SCHUBERT, ENGINEERING THE CLIMATE. SCIENCE, POLITICS, AND VISIONS OF CONTROL (2021).

[93]*Supra* B.

[94]Lakens, *supra* note 45 (reaching the same conclusion).

because the discussed piece of evidence may be flawed, this certainly does not mean that judges cannot be biased or that computer-based decision-aids cannot be put to socially beneficial use. But when it comes to human flaws and biases in judging, more collaborative research is needed.[95] This, in turn, would lead to a better assessment of where AI can actually be of help.

---

[95]*See also* Edmond & Martire, *supra* note 2, at 664 ("If justice should not only be seen to be done but, more fundamentally, be done, then judges and scientists may need to begin a slow process of dialogue and research in order to better understand and perhaps reform the way that our complex societies organise legal practice and decision-making.").