What the Judge Argues is Not What the Judge Thinks

Eye Tracking Evidence about the Normative Weight of Conflicting Concerns in a Torts Case

Christoph Engel
Rima Maria Rahal
What the Judge Argues is Not What the Judge Thinks
Eye Tracking Evidence about the Normative Weight of Conflicting Concerns in a Torts Case

Christoph Engel & Rima Maria Rahal

January 2020
revised March 2021
What the Judge Argues is Not What the Judge Thinks
Eye Tracking Evidence about the Normative Weight of Conflicting Concerns in a Torts Case

Christoph Engel & Rima Maria Rahal

Abstract

Judicial decision making is not a mechanical activity. It requires a voluntary act. In the abstract, the judge must strike a balance between incompatible normative goals, like backward looking compensation and forward looking deterrence. In the concrete, the decision-maker must weigh conflicting facets of the conflict of life. As a rule, judicial decisions come with explicit reasons. These reasons rationalize the decision. In this paper, we use eye tracking as a window into the – consciously not fully accessible – process of making the decision, testing laypersons on a run-of-the-mill torts case. We manipulate the degree of normative conflict, and either have participants decide as judges, or plead as attorneys. Against expectations, attention is not chiefly directed towards items that support the final outcome. Rather outcome is predicted by attention on potentially conflicting items. Explicit reasons and fixations on items are essentially uncorrelated. Decision-makers are not aware of the elements of the evidence that have been critical for their decision.

JEL: D01, D81, D91, K13, K40

Keywords: legal decision making, context of discovery vs. context of representation, eye tracking, balancing, torts, compensation, deterrence
I. Introduction

Judges are no machines. They do not calculate the outcomes of legal cases, they decide. This voluntary act of a disinterested person is necessary since the typical legal case is ill-defined. In this paper we bracket ambiguity about the facts (on that see Glöckner and Engel 2013a, Engel, Timme et al. 2020), and focus on the ambiguity of the law. Experimental participants are confronted with an ambiguous case. While some features of the case support the claim, others speak against it. The features belong to different normative domains. On conceptual grounds, they are incompatible. Deciding the case requires balancing (cf. Aleinikoff 1987). There is a long-standing normative debate at the meta level about the desirability of balancing (Pildes 1993, Alexy 2003, Petersen 2017). Yet, for legal decision-making in practice, there is no way around. We are interested in how decision-makers actually weigh the conflicting features.

As the outcome of balancing is not fully predictable, the law goes a long way to secure impartiality. As a rule, judges are required to justify their decisions, usually even in writing (for the normative debate see Engel 2007). The competent judge must have been determined before the case reaches the court. If an outsider might suspect the judge to be biased, she is recused. If one of the parties worries that the judge has not properly assessed the normative weight of competing concerns, she may appeal the case, to list only the most important safeguards. Still, after the fact, other authorities can only reconstruct how the competent judge has balanced the competing features of the case, or one can have a new judge redo the task.

In the past, there was no way to go further. In the middle of the previous century, Bertold Brecht could still write: “but the thoughts are free”. No longer. With eye tracking, this generation has a chance to see choices in the making. We exploit the fact that there is a close relationship between eye gaze and attention (for a discussion of the potential and limitations of the method see Just and Carpenter 1980) to catch a glimpse of the legal mind in action.

To that end, we ask experimental participants to decide a deliberately ambiguous case. While preparing their decision, participants see a screen on which the competing concerns, and the corresponding features of the case, are presented. We record the number of fixations on each of the eight items, and their duration. After they have decided, we ask participants to rate the importance each item has had for their decision, the perceived degree of conflict between the item in question and other items, the difficulty they had in assessing the item, and the degree of ambiguity they have perceived. This set of measures makes it possible to compare the conscious assessment with data about the actual decision-making process. We find a pronounced difference between statements and gaze data. We can show how the context of discovery and the context of representation fall apart.

We had expected a closer mapping between eye gaze and the normative evaluation of a feature that favours plaintiff or defendant, respectively. This is why we have added two experimental manipulations. We compare two versions of our torts case. In one version, normative conflict is pronounced. In this version, a backward-looking perspective would call
for full compensation, while a forward-looking perspective would favour partial compensation only, to avoid that a small amount of negligence causes defendant to be bankrupt. In the other version of the case, this element is absent. There is still normative conflict. But it is rather likely that a judge would decide for plaintiff. Our second manipulation introduces a party perspective. Using the version of the case with the serious normative conflict, we also test participants in the role of representative for plaintiff or representative for defendant.

Our manipulations matter for outcomes. In the conflict case, mock judges are more likely to only partly compensate plaintiff. The plaintiff’s representatives ask for more money than defendant’s representatives are willing to give. These differences are largely reflected in the stated weight a participant assigns to one of the conflicting features of the case. Yet against our expectations, there is no mapping between the assigned role and the number or the duration of fixations on the items on the screen.

Eye gaze does, however, have very high predictive power for the decision. Knowing the number or the duration of fixations on the conflicting features of the case makes it possible to predict the decision with impressive precision. Interestingly, a high number or duration of fixations on the items that favour the defendant predicts a higher, not a lower verdict. The eye data suggests that it matters how strongly the participant has engaged with potential arguments that speak against the decision she ultimately makes. We can even use the evidence to construct a counterfactual. If the two defense items on which judges are most likely to fixate had not been included, our regressions predict a verdict that would have favoured the defense.

We can even go a step further, and simultaneously explain the decision by the stated weight for a feature of the case, and the number or duration of fixations on it. Eye gaze data not only remains predictive. It predicts the decision even better than the stated reasons. Again the effect of items favouring defense is most interesting. While the stated weight of these items is the lower the higher the verdict (and therefore the more in favour of plaintiff), also in these regressions the number and the duration of fixations tell the opposite story. Participants seem to not be aware that they engage with the counterarguments before deciding in favour of plaintiff.

We thus observe a pronounced difference between the context of discovery and the context of representation. This distinction has originally been developed in analytical philosophy (Popper 1935: 113, Ziman 2000). Generating a decision and representing it to its respective audience(s) are independent activities (Luhmann 1966). The normative debate about representation norms tends to consider this insight from a public choice perspective. One is concerned that biased decision-makers might hide normatively unacceptable motives behind seemingly innocent justifications (Schauer 1995). We are not arguing that this risk should be neglected. Yet our data suggests that the distinction even matters if the decision-maker is perfectly good-natured, and exclusively motivated to implement the intentions of the law. The actual decision-making process may not be available to introspection (Nisbett and Wilson 1977, Glöckner and Engel 2013a, Engel, Timme et al. 2020), as cognitions about
the thought process may be overlaid with post-hoc rationalizations or other interferences of retrieval and reporting.

The normative implications of this insight are profound. If the law is concerned about the quality of judicial decisions and their impartiality, it is not enough to combat the deliberate miscarriage of judicial decision-making. Our finding can only be the start of a line of research that, step-by-step, uncovers the likely determinants of judicial decision-making. It will be particularly important to understand in which ways these mental mechanisms can be manipulated by interested outsiders. After all, in the adversarial system such attempts are even institutionalized.

II. Design

In our experiment, we give participants a stylized torts case. The case is deliberately ambiguous. The case comprises four elements in support of plaintiff’s claim: defendant has caused damage to plaintiff’s property; defendant has violated a rule that one might argue also protects plaintiff’s property; damage is pecuniary, so that plaintiff can be compensated; if the court decides for plaintiff, others will be deterred from committing comparable torts. Yet the case also comprises four elements that might favor defense. Plaintiff has herself put her property at risk; defendant has acted with good intentions; the obligation to pay damages might hit defendant severely; it might have a chilling effect on third parties. These last two elements are manipulated. In the no conflict condition, the amount of damage is moderate, so that these concerns are rather far-fetched. By contrast in the conflict condition, defendant would have to file for bankruptcy when obliged to fully compensate plaintiff.

The conflict manipulation is motivated by the debate about the normative purpose of torts liability (for an indepth discussion see Schwartz 1994, Schwartz 1996, Oberdiek 2008, Oberdiek 2014). The most salient distinction is the one between compensation and deterrence (Geistfeld 2011). From a compensation perspective, damages are correlates of property. If a good is A’s property and not B’s, it is for A to decide what to do with the thing. If B impinges upon A’s sphere of influence, and this makes the good less valuable for A (or even destroys it altogether), A must be made whole (more from Geistfeld 2014, Goldberg, Sebok et al. 2016). Compensation is an exercise of restorative, distributive justice (Shuman 1993, Gardner 2013, Sheinman 2014). The law steps in after another person’s property has been damaged. From this perspective, the conflict manipulation is immaterial. How severely defendant is hit by the obligation to pay is irrelevant for the restorative purpose of the rule. If at all, the fact that the damage is large makes it even more important that plaintiff is compensated. She has no responsibility for defendant negligently inflicting harm on her.

Alternatively, policy makers may however aim at prevention, rather than reparation. In this forward-looking perspective, the prospect of having to pay damages deters rule violations (Eisenberg and Engel 2014, Eisenberg and Engel 2016). When factoring in the payment of
damages, in expectation the tortfeasor hurts herself by violating the rule of non-inter- 
vention. Realizing that rule violations do not pay, she desists from the socially harmful activity
(Landes and Posner 1987). In this perspective, the conflict manipulation matters. The obliga-
tion to pay works as a sanction. Forcing defendant to file for bankruptcy is not only ex-
cessive. It may even backfire. Defendant may be induced to gamble, as she could anyhow 
not continue her previous life when held liable. This argument has for instance been ad-
vanced in favor of capping the amount to be paid in the case of medical malpractice
(Eisenberg and Wells 2006, Matsa 2007). The risk of liability may loom so large on individ-
uals’ minds that they desist from socially acceptable activities. If they overreact, there might 
be a socially harmful chilling effect (De Geest 2012).

In a concrete torts case, not only the normative foundations of the remedy may be at stake. 
It may also be debated whether it suffices for plaintiff to show that defendant has harmed 
her, or whether some action or omission of plaintiff constitutes contributory or comparative 
negligence (for background see Bohlen 1907, James Jr 1952, De Mot 2013). In principle, 
the negligence standard is an objective one. Plaintiff must show that defendant has violated 
a duty of care. Yet some authors have argued for a subjective element (Ben-Shahar and 
Porat 2016), which would allow defendant to advance an excuse (Goldberg 2015). Acting 
with good intentions might be one such excuse.

Participants are assigned the role of judge, counsel for plaintiff or counsel for defendant 
and read the following vignette:

A earns 3,000 € / month. He does not own any property that could be seized. In his 
car, he passes through a village. The speed limit is 30 km/hour. A drives at 55 
km/hour. A cat jumps onto the street. A breaks to save the cat. He loses control of 
the car. The car collides with an aboveground tank owned by B. Since the 
tank had been put in place, construction law has changed. Today it would no longer 
be permissible to build the tank aboveground. An expert testifies that A would not 
have lost control of the car had he obeyed the speed limit. A’s insurance does not 
cover damage inflicted on aboveground objects off the street.

Repairing the tank costs 1,500 €. There is a risk of groundwater contamination. This 
is why the police obliges B to dredge the ground, at the cost of 43,500 €. B sues A, 
and asks for 1,500 € 45,000 €.

Plaintiff demands full compensation. Defendant argues that at most partial compen-
sation (500 € 15,000 €) would be justified.

Treatment variations are colour-coded; in the experiment, the manipulated items have of 
course not been highlighted. In the judge conditions, participants are asked to decide the 

case. They know that they will also be asked to give a written justification. In the plaintiff

1 Under German law, the owner of land can be held responsible if an event on her land puts groundwater 
at risk. The police is free to ask the owner to take care of the risk, even if the event has been caused by 
a third party. In that case, it is left to contract or torts liability to indemnify the owner for the cost of 
obliging with the police order.
and defendant conditions, they are asked to indicate how the court should decide, and to plead in favour of their respective client.

In preparation of their decision, participants see the computer screen reproduced in Figure 1. This decision screen is in the spirit of a decision aid. It makes an offer for structuring the normative thinking about the case at hand. There is one column with features of the case supporting plaintiff, and another column with features supporting defendant. The position of the two columns is counterbalanced between subjects. The four rows cover facts regarding the responsibility of plaintiff and defendant, the remedy, and potential effects on third parties. Rows are counterbalanced between subjects as well. Participants know that, while they look at this screen, there eye movements are recorded.

<table>
<thead>
<tr>
<th>Kläger</th>
<th>Beklagter</th>
</tr>
</thead>
<tbody>
<tr>
<td>tat nichts</td>
<td>oberirdisch</td>
</tr>
<tr>
<td>zu schnell</td>
<td>rettete Katze</td>
</tr>
<tr>
<td>wieder-hergestellt</td>
<td>geht bankrott</td>
</tr>
<tr>
<td>vor-sichtiger</td>
<td>abgeschreckt</td>
</tr>
</tbody>
</table>

**Figure 1**
Decision Screen

middle column: items in support of plaintiff’s claim
(above referred to as causation, negligence, compensation, deterrence)

righthand column: items in support of defense
(above referred to as contributory [negligence], excuse, excessive [deterrence], chilling)

For an English translation, see Appendix

On the next screen, participants are asked to spell out the reasons for their decision or claim. They can write up to one page. On the following screen they are asked for the amount of money they want defendant to pay. Thereafter participants are reminded of the eight features of the case that had been presented on the screen used for tracking their gaze patterns. For each of these concerns, they are asked to indicate, on a Likert scale from 1 (not at all) to 7 (extremely so), how much weight the argument should carry; how difficult they found it to understand the argument; to which degree the argument was ambiguous; to which degree the argument was in conflict with other arguments. Each set of questions is presented to them on a separate screen.²

² For exploratory purposes, we had also asked participants to rate the weight they would assign to nine abstract normative concerns that might be related to the case; we measured social value orientation
The experiment was run in the computer lab of the Bonn Max Planck Institute. 139 students participated in the experiment, 36 in the judge no conflict condition, 34 in the judge conflict condition, 35 in the role of plaintiff, and 34 in the role of defendant. One of the eye trackers did not properly isolate fixations. For the analysis, we only use data from the 116 participants where eye gaze has been properly recorded (30 in the judge no conflict condition, 29 in the judge conflict condition, 29 in the plaintiff condition, at 28 in the defendant condition). 71 of these participants indicated their gender as female, 43 as male, 1 as diverse, and 1 did not respond to this question. Participants earned an average of 15€ per hour for their participation.

Eye gaze was recorded with binocular remote Eye Tribe trackers at a 60Hz sampling rate following a 9-point calibration procedure. Participants were seated at about 60cm distance from 14” Dell Latitude E5440 laptops with a native resolution of 1366 x 768 pixels, and eye trackers were mounted directly below the screen. We defined two types of areas of interests (AOI) on the grid screens, choosing to maximize AOI size instead of using the small preregistered AOI sizes. Non-target AOIs containing labels were defined as 190 x 170 and 280 x 170 pixels in size. Target AOIs containing arguments associated with each category are defined as 280 x 340 pixels in size. We defined fixations with a 30 pixel tolerance in the summed deviation of points’ maximum and minimum coordinates on the x- and y-axes and a minimum duration of 50ms (Salvucci and Goldberg 2000).

III. Hypotheses

The main innovation of our paper is the use of eye tracking as a window into the decision of a legal case in the making. This is why our hypotheses focus on the effect of the experimental manipulations on the decision-making process. We rely on the assumption that attention and eye movements are linked (Just and Carpenter 1980), especially in complex tasks (Rayner 1998), to the extent that “the most active location in working memory will eventually determine the most likely direction of the eye movement at a given point in time” (Huettig, Olivers et al. 2011: 141). We expect that the attention participants pay to a specific normative concern is reflected in the number of fixations to the associated area of interest (AOI) on the screen (Orquin and Loose 2013), and in the total duration of fixations on the AOI (Rayner 1998). Specifically, we expect that cases with normative conflict require participants to access more information than cases without conflict (number of fixations, Fiedler, Glöckner et al. 2012), and that they need more time to process the information (duration of

(Murphy and Ackermann 2014); justice sensitivity via the Justice Sensitivity Inventory (Schmitt, Baumert et al. 2010); personality via the 24-item Brief HEXACO Inventory (De Vries 2013), intuitive vs. rational decision making style via the Rational and Intuitive Decision Styles Scale (Hamilton, Shih et al. 2016) and third-party inequality-inefficiency trade-off (Rahal, Hoeft et al. 2020).

As we do not need these supplementary measures for the message of the paper, we do not report them. The supplementary data is available at https://osf.io/nu4qj/?view_only=8948b4037f8941dd9f00f0c236bbf1c5.

3 We have preregistered the hypotheses at https://osf.io/nu4qj/?view_only=8948b4037f8941dd9f00f0c236bbf1c5.
fixations). Recall that the conflict manipulation is confined to the “remedy” and “other” items (last two rows in Figure 1). We predict

**Hypothesis 1 (conflict)**

a) In the conflict cases, total response time, the duration of fixations on content items, and the number of fixations on these items are higher than in the no-conflict cases.
b) In the conflict cases, the duration of fixations and the number of fixations on the “remedy” and “other” items is higher than in the no-conflict cases.
c) In the conflict cases, the duration of fixations and the number of fixations on the “remedy” and “other” items is higher than on the remaining items.

In the adversarial system, the counsel for plaintiff and the counsel for defendant are not only allowed, they are even supposed to be partisan. In the interest of creating a level playing field, either party to the case may expect her counsel to present her respective cause in the most favourable light. The manipulation thus introduces an institutionalised bias. In the logic of the reasons motivating the hypotheses for judges we expect this bias to translate into the way how participants in the role of counsel process the information. We therefore predict:

**Hypothesis 2 (counsel)**

a) The number and the duration of fixations allocated to the content AOIs in the “client” column is higher than to the content AOIs in the “opponent” column.
b) The pattern of fixations differs between attorneys for the plaintiff and for the defendant.
c) Attorneys exhibit a lower total response time, a smaller number of fixations, and a lower fixation duration, than judges seeing the same case.
d) The pattern of fixations is more balanced in judges (seeing the same case) than in attorneys.

### IV. Results

#### 1. Manipulation Check

Figure 2 shows that our manipulations have worked. The disputed amount in the judge no conflict condition is only 1/30 of the amount that is at stake in the conflict conditions. Therefore, for comparability the figure (and all later analysis) normalises the ruling or claim to a fraction of full compensation. In the judge no conflict condition, on average 80.22% of the maximum amount is awarded. 53% of the judges award the full amount. By contrast, if there is a conflict between compensation and deterrence, judges on average only grant 41.71% of the maximum. Only 14% award the full amount. The modal award is what defendant is

---

4 We also had preregistered the following hypothesis regarding the relationship between eye gaze and self-report: “The higher the importance rating of arguments against the client in the rating task, the smaller the difference in the proportion of attention allocated to the content AOIs in the “client” vs. “opponent” columns”. We discuss this relationship below 4.
happy to grant, i.e., a third of total damage. The difference in awards between the judge no conflict and the judge conflict condition is highly significant (Mann Whitney, N = 59, p < .001).

The latter finding stands in stark contrast to the claims from experimental counsels for plaintiff. They on average ask for 91.11% of the maximum amount. 68% of the counsels ask for the full amount. The difference in outcomes between judges and counsels for plaintiff is also highly significant (Mann Whitney, N = 57, p < .001). Counsels for defendant do also behave as expected. They on average only concede 30.16% of total damage. 86% do only concede a third, or even less of total damage. Descriptively, this is even less than judges grant when faced with the normative conflict. Yet the difference between the verdicts of judges and the pleadings of counsels for defendant is not significantly different from zero. With this slight qualification we conclude that both the conflict and the role manipulations have worked as designed.

![Figure 2](Verdict by Treatment)
2. Effect of Experimental Manipulations on Eye Gaze

We had expected that the presence of normative conflict would make it harder for judges to decide, and that this would lead to higher response time, a higher number of fixations on all content items, and a longer duration of fixations on these items (H1a). Yet none of these expectations is borne out by the data. We do not find any evidence for treatment effects on these dependent variables, neither with non-parametric (Table S1, Model 1) nor with parametric tests (Table S2, Model 1).

We had further expected that the presence of normative conflict would induce judges to fixate more frequently on the critical items (regarding remedy and the effect on third parties), and to spend more time on these items (H1b). Again, the expectations are not supported by the data. For these dependent variables, we do not find any significant treatment differences, neither non-parametrically (Table S1, Model 2) or parametrically (Table S1, Model 2).

We had also expected that participants, in the conflict condition, would fixate more frequently, and for longer time, on items that characterise the conflict, i.e. on the remedy items and the items describing potential effects on third parties (H1c). Yet once more we do not find a significant difference, neither non-parametrically (Table S1, Model 3) nor parametrically (Table S2, Model 3).

We also do not find support for the hypotheses regarding the role manipulation. We had expected that counsels for defendant would fixate more frequently on items supporting defendant’s cause (Table S3, Model 1, and Table S4, Model 1), and that counsels for plaintiff would fixate more frequently on items supporting plaintiff’s cause (Table S3, Model 2, and Table S4, Model 2) but there is no evidence for such a pattern, whether we use non-parametric or parametric tests (H2a). In the same vein, we found no evidence that counsels for defendant spent more time fixating on items supporting defendant’s case (Table S5, Model 1, and Table S6, Model 1), or that counsels for plaintiff spent more time fixating on items supporting plaintiff’s cause (Table S5, Model 2, and Table S6, Model 2, H2a).

We also found no evidence in support of the expectation that counsels for plaintiff fixate more frequently on items favouring plaintiff than counsels for defendant, or that counsels for plaintiff spend more time on items favouring plaintiff than counsels for defendant (H2b, Table S7).

Judges have to strike a balance between the competing normative concerns, while counsels for either plaintiff or defendant just have to argue for their client’s interest (H2c). This is why we had expected that judges need more time overall and show longer fixations (Table S8, Model 1 and Model 3), which is not supported by the data. We do, however, find a significant difference in the total number of fixations on these items (Table S8, Model 2). Judges on average fixate on 44.48 items, while defendants on average fixate on 16.62 more (p = .067), and plaintiffs fixate on 15.69 more (p = .081). Yet these effects point into the opposite direction of what we had expected (H2c). If we consider the relative amount, or the relative duration, of fixations on items favouring plaintiff (or defendant for that matter), we
do not find a significant difference between judges and counsels (Table S9). Hence, against our expectations ($H_2d$), there is no evidence for a more balanced fixation pattern in judges than in counsels.

3. Predictive Power of Eye Gaze

One might therefore think that eye tracking is simply not helpful for understanding decision-making in a legal context. In this section we show that this conclusion would be wrong. As the regressions in Table 1 demonstrate, with only information about the number or the duration of fixations, and the condition, we can predict the amount of damages that the participant assigns, requests, or is willing to grant. We can do so with surprisingly high precision. The adjusted $R^2$ of the regressions in Table 1 is between .819 and .878. In all regressions, the root mean square error is fairly low (between .226 and .299). Figure 3 visually represents the predictive power of gaze data, by comparing observed choices with the choices predicted when only using gaze data, and treatment. In Table 4 in the Appendix, we show that the predictive power of eye gaze even remains very high even if we conceal treatment information.

A closer look at coefficients and significance levels further shows that eye gaze is not only predictive. We learn which items have carried most weight in the decision made by the participant in question. Some of the significant effects are intuitive: the more frequently, and the longer, a participant has focused on the causation and compensation items, the more the verdict favours plaintiff. In several specifications, we also find a related effect for fixations on the deterrence item. Yet in all specifications, we also find significant effects of the contributory negligence item, and on the items evoking the risk of defendant becoming bankrupt, and of a chilling effect on third parties. Now all these coefficients are positive. The more gaze data shows that the participant has engaged with these items favouring defense, the more she decides in favour of plaintiff. This suggests that an important element in the decision-making process is elaborating on potential counterarguments against the decision the participant plans to make.

---

5 Our explanatory variables are the fraction of all fixations, or the fraction of total duration on one of these 8 content items, on the item in question. Hence by design for each participant the 8 fractions add up to 1. Were we to estimate a regression with a constant, one of the items would drop out. Coefficients would no longer inform us about the relative weight of the item in question for the decision, but about the difference in weight between the omitted item and the remaining items. This is why we estimate models without a constant.

6 The maximum of course being 1.

7 In the paper, we focus on the regressions with treatment information as this facilitates interpretation. Coefficients and significance levels on all regressions look similar if we include the data from the 23 participants where we must be sceptical about the performance of the eye tracking device (see Table S10).
## Table 1
### Predictive Power of Eye Gaze

<table>
<thead>
<tr>
<th></th>
<th>Number of fixations</th>
<th>Fixation Duration</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Model 1</td>
<td>Model 2</td>
</tr>
<tr>
<td></td>
<td>Conflict manipulation</td>
<td>Role manipulation</td>
</tr>
<tr>
<td>causation</td>
<td>0.998** (.306)</td>
<td>0.909*** (208)</td>
</tr>
<tr>
<td>negligence</td>
<td>.285 (.557)</td>
<td>.272 (.312)</td>
</tr>
<tr>
<td>compensation</td>
<td>1.069** (.372)</td>
<td>0.385* (.172)</td>
</tr>
<tr>
<td>deterrence</td>
<td>.836* (.377)</td>
<td>.464 (.278)</td>
</tr>
<tr>
<td>contributory</td>
<td>.827** (.251)</td>
<td>.486** (.136)</td>
</tr>
<tr>
<td>excuse</td>
<td>.623 (.407)</td>
<td>-.207 (.238)</td>
</tr>
<tr>
<td>excessive</td>
<td>1.006** (.304)</td>
<td>.505** (.183)</td>
</tr>
<tr>
<td>chilling</td>
<td>.615* (.264)</td>
<td>.421** (.135)</td>
</tr>
<tr>
<td>conflict</td>
<td>-.401*** (.083)</td>
<td>-.301*** (.079)</td>
</tr>
<tr>
<td>plaintiff</td>
<td>.484*** (.064)</td>
<td></td>
</tr>
<tr>
<td>defendant</td>
<td>-.163* (.064)</td>
<td>-.163* (.064)</td>
</tr>
<tr>
<td>N</td>
<td>59</td>
<td>85</td>
</tr>
<tr>
<td>RMSE</td>
<td>.299</td>
<td>.230</td>
</tr>
<tr>
<td>adj. R²</td>
<td>.819</td>
<td>.873</td>
</tr>
</tbody>
</table>

**OLS causation, negligence, compensation, deterrence, contributory, excuse, excessive, chilling: shorthands for items on the screen of Figure 1 all items represented with proportion, per participant, of fixations on all 8 items / duration of these fixations models without constants, as proportions add up to 1 by design conflict: excessive and chilling items induce normative conflict plaintiff, defendant: assigned role; reference category in this comparison: judge conflict one (aggregate) datapoint per participant standard errors in parenthesis

*** p < .001, ** p < .01, * p < .05, + p < .1
Ultimately with our data we can of course not prove that participants have focused more frequently, and for a longer time, on items opposed to the decision they were about to make, because this was necessary to neutralise potential counterarguments. But if we trust that the regressions not only organise the data, but make a prediction at the population level, we can back the interpretation with out of sample predictions. Which outcome would the regression predict, for each participant, if one or even two of the items supporting defense had not been made? This is a meaningful question as it takes sufficient knowledge of the law, and sufficient litigation expertise, to act as counsel for one of the parties. It is therefore not unlikely that some parties have to live with less competent attorneys. What would be the effect on the ruling if some of the arguments supporting defense had not being introduced?

The left panel of Figure 4 shows how, according to Model 1 of Table 1, judges would decide if the most predictive argument in favor of defense is not mentioned, the one pointing out that defendant might be forced to file for bankruptcy. The right panel shows the predicted rulings if, additionally, the second most predictive argument is dropped, arguing that there is contributory negligence. As one sees, having a chance to balance out the argument, but ultimately deciding against it, indeed favors plaintiff. If predictions were on the 45° line,

---

8 Predictions are calculated the following way: the number of fixations on the items in question is set to 0. The fraction of the remaining items is newly calculated. These recalculated fractions are substituted into the regression, to generate the predicted ruling. The picture looks very similar when using the duration, rather than the number of fixations (Figure S1).
dropping the argument in question would be immaterial. The predicted decision would not differ from the observed decision. Yet as one sees, in the left panel only three decisions are on the line. All other predictions are below the line, i.e. would favor defendant. Dropping the second argument in favor of the defendant further reduces predicted rulings.

![Figure 4](image.png)

**Figure 4**
Out of Sample Prediction When Dropping Defense Items

4. Interpretation of Eye Gaze

When designing the experiment, we had expected that having participants rate the eight items characterizing the competing normative concerns would help with the interpretation of the number or the duration of fixations. This is why we asked them to rate, on a 7 point Likert scale, the weight they assign to each of the items; the degree of conflict they perceive between the item in question and other items; the difficulty they have in understanding the item; the degree of ambiguity they perceive. Yet the regressions in Table 2 show no evidence that these self-report measures and eye data are related.
In retrospect, this null result is not that surprising. It has been shown before that the mental process that enables human decision-makers to dissolve ambiguity cannot be modelled as an algebra-like process (Pennington and Hastie 1986). Individuals are engaged in sense-making, and constructing coherent interpretations (or stories) from, the available evidence (e.g., Pennington and Hastie 1992, Holyoak and Simon 1999, Robbenbott, Darley et al. 2003, Thagard 2006). It has been argued that the process of coherence construction can be formally described by parallel constraint satisfaction (PCS) network models (e.g., Thagard 1989, Kunda and Thagard 1996, Read, Vanman et al. 1997, Simon and Holyoak 2002, Robbenbott, MacCoun et al. 2010, Jekel, Glöckner et al. 2018). These models have also been successfully applied to legal decision-making (e.g., Holyoak and Simon 1999, Carlson and Russo 2001, Simon, Snow et al. 2004, Engel and Glöckner 2013, Glöckner and Engel 2013b, Simon, Stenstrom et al. 2015). Critically for the present context, the mental process is subconscious; only the final outcome, i.e. the interpretation, is propelled back to conscience (Glöckner and Betsch 2008). If participants, after having decided, assign a high weight to an item, this indicates that they rationalize the decision with the weight, not that the item has had the biggest impact when actually deciding. The weights, and the remaining indicators, are given in the context of representation; eye gaze is a window into the context of discovery.

<table>
<thead>
<tr>
<th></th>
<th>model 1</th>
<th>model 2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>number of fixations</td>
<td>duration of fixations</td>
</tr>
<tr>
<td>weight</td>
<td>-.0018 (.0024)</td>
<td>-.0021 (.0026)</td>
</tr>
<tr>
<td>conflict</td>
<td>.0026 (.0024)</td>
<td>.0026 (.0026)</td>
</tr>
<tr>
<td>difficulty</td>
<td>-.0015 (.0037)</td>
<td>-.0013 (.0041)</td>
</tr>
<tr>
<td>ambiguity</td>
<td>.0029 (.0034)</td>
<td>.0028 (.0038)</td>
</tr>
<tr>
<td>cons</td>
<td>.1208*** (.0137)</td>
<td>.1221*** (.0151)</td>
</tr>
<tr>
<td>N obs</td>
<td>928</td>
<td>928</td>
</tr>
<tr>
<td>N uid</td>
<td>116</td>
<td>116</td>
</tr>
</tbody>
</table>

### Table 2
**Interpretation of Eye Gaze**

- **dvs**: relative number of fixations and relative duration of fixations on one of 8 content items
- **ivs**: stated weight, degree of conflict, difficulty or ambiguity of item in question
- **standard errors in parenthesis**
- ***** p < .001**
5. Comparing the Predictive Power of Self-Report vs. Gaze Patterns

In the final analytic step, we add the stated weight participants assign to each of the eight items on the decision screen to the statistical models of Table 1.\(^9\) Comparing the adjusted R\(^2\) between the corresponding models of Table 3, we learn that the explanatory power only increases very little. In the models focusing on the treatments involving normative conflict, the increase in explanatory power is minuscule (from .875 to .886 in model 2, and from .877 to .890 in model 4). In the models focusing on judicial decisions, the increase is slightly bigger (from .821 to .894 in model 1, and from .825 to .895 in model 3), but still fairly modest. With very few exceptions, coefficients of gaze data that were significant when not controlling for reported weight remain significant when adding reported weights as controls.\(^10\)

This strengthens the claim that the mental process leading to the decision, and their ex-post rationalization, are at least partly disjunct. Actually, for the only weight coefficient that is (in all four specifications) significant at conventional levels (contributory), the corresponding measure for the relative number or the relative duration of fixations is significant as well, but has opposite sign. The more a participant argues that the possibility of contributory negligence is an important concern, the lower her verdict. Yet the more she has visually focused on this item, the higher her verdict.

<table>
<thead>
<tr>
<th></th>
<th>number of fixations</th>
<th>fixation duration</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>conflict manipulation</td>
<td>role manipulation</td>
</tr>
<tr>
<td>weight causation</td>
<td>.030 (.020)</td>
<td>.014 (.018)</td>
</tr>
<tr>
<td></td>
<td>(.020)</td>
<td>(.018)</td>
</tr>
<tr>
<td>weight negligence</td>
<td>-.005 (.026)</td>
<td>.010 (.024)</td>
</tr>
<tr>
<td></td>
<td>(.024)</td>
<td>(.026)</td>
</tr>
<tr>
<td>weight compensation</td>
<td>.052* (.027)</td>
<td>.023 (.019)</td>
</tr>
<tr>
<td></td>
<td>(.019)</td>
<td>(.028)</td>
</tr>
<tr>
<td>weight deterrence</td>
<td>.030 (.021)</td>
<td>.011 (.017)</td>
</tr>
<tr>
<td></td>
<td>(.021)</td>
<td>(.017)</td>
</tr>
<tr>
<td>weight contributory</td>
<td>-.081*** (.021)</td>
<td>-.035* (.017)</td>
</tr>
<tr>
<td></td>
<td>(.021)</td>
<td>(.017)</td>
</tr>
<tr>
<td>weight excuse</td>
<td>-.003 (.021)</td>
<td>.0003 (.017)</td>
</tr>
<tr>
<td></td>
<td>(.021)</td>
<td>(.017)</td>
</tr>
<tr>
<td>weight excessive</td>
<td>-.038* (.022)</td>
<td>-.027 (.016)</td>
</tr>
<tr>
<td></td>
<td>(.022)</td>
<td>(.016)</td>
</tr>
<tr>
<td>weight chilling</td>
<td>.009 (.024)</td>
<td>-.002 (.020)</td>
</tr>
<tr>
<td></td>
<td>(.024)</td>
<td>(.020)</td>
</tr>
<tr>
<td>fixations causation</td>
<td>.707* (.340)</td>
<td>.842** (.312)</td>
</tr>
<tr>
<td></td>
<td>(.340)</td>
<td>(.312)</td>
</tr>
</tbody>
</table>

\(^9\) Results look similar, and in particular the effect of gaze data remains significant, if we replace the stated weight of the item in question with the stated degree of conflict (Table S11), the stated difficulty in assessing the item (Table S12), or with the perceived degree of ambiguity (Table S13). All of these self-report scores have even less explanatory power.

\(^{10}\) In the regressions of Table 3, fixations on the compensation item are insignificant, while they are significant in the corresponding models 2-4 in Table 1. In Table 3 fixations on the chilling item only have a weakly significant effect, while the effect is significant at conventional levels in the corresponding models of Table 1.
| fixations | negligence | .364 (.340) | .303 (.361) | .298 (.554) | .06 (.367) |
| fixations | compensation | .459 (.409) | .303 (.361) | .586* (.335) | .365 (.238) |
| fixations | deterrence | 1.266** (.373) | .557 (.351) | .979** (.336) | .612* (.312) |
| fixations | contributory | .809* (.316) | .494* (.238) | .884** (.320) | .481* (.229) |
| fixations | excuse | .589 (.394) | -.095 (.295) | .355 (.440) | -.259 (.338) |
| fixations | excessive | .939* (.359) | .513* (.290) | .925** (.296) | .525* (.255) |
| fixations | chilling | .529* (.271) | .4933* (.207) | .628* (.258) | .485* (.195) |
| conflict | plaintiff | -.256** (.069) | .453*** (.066) | .449*** (.066) |
| conflict | defendant | .123* (.064) | -.127* (.063) | |
| N | 59 | 85 | 59 | 85 |
| RMSE | .225 | .220 | .226 | .214 |
| adj.R² | .897 | .884 | .897 | .891 |

Table 3
Reported vs. Observed Impact of Items on Verdict

V. Discussion

Human thought processes are powerful. Yet human thought processes can be on the wrong track. As this paper shows, this may also hold for researchers working on human thought processes. In our experiment, we have exposed lay participants to a deliberately ambiguous case. The ambiguity has made it impossible for participants to derive the decision from a single normative principle by way of deductive reasoning. We have forced them to use their intuition. In the spirit of causal identification, we have randomly assigned participants to treatments. In one dimension we have manipulated the degree of normative conflict, and hence the degree of challenge when deciding the case regardless. In the other dimension we have manipulated bias. We have compared a judge instructed to find an appropriate decision with a counsel for either plaintiff or defendant, instructed to plead in favour of their assigned client. We have designed the experiment in the expectation that the importance
of one of eight features of the case for fulfilling the task would be reflected in the number and in the duration of fixations on the respective items. Hence, we had expected that the items that aggravate the degree of normative conflict in the conflict manipulation would attract attention, which we hypothesized to be reflected in fixations. Likewise, we had expected that representatives of one of the sides of the dispute would predominantly have their eyes dwell on the items that support their respective cause. This is not what we have found. The manipulations do not significantly explain the number or the duration of fixations.

In the interest of disentangling alternative interpretations for fixations, we have asked participants to rate the weight they attach to each of the eight items; the degree of conflict between the item of question and alternative items on the screen that they perceive; the difficulty they experience in understanding the importance of the respective item; the degree of ambiguity they perceive with respect to the item. Again, against our expectations we do not find any significant correlation between these self-report measures and the number or the duration of fixations on the respective item.

Yet digging deeper into our data we find that, when only using gaze data, we can predict choices with surprising precision. The predictive power is even higher if we control for treatment. If we add any of the self-report measures, the accuracy of the prediction increases only so slightly. More tellingly even, to the degree that the reported weight on an item and either the number or the duration of fixations on the item both turn out significant, coefficients have opposite sign. If self-reported weight suggests a decrease in the verdict, gaze data suggests an increase. Specifically, the significant positive coefficients of fixations on items favouring defence indicate that decision-makers must elaborate on potential counterarguments before deciding against them.

Our findings can be rationalised with a long-standing claim from decision research: The process of constructing a choice is difficult to introspectively assess and report (Nisbett and Wilson 1977). Often, only the final outcome is reported (Glöckner and Betsch 2008). Explaining the outcome is not a mirror image of generating it. The context of discovery and the context of representation are distinct processes.

The legal order rightly cares about the context of representation. It helps the parties accept the ruling if it comes with a consistent explanation. The explanation may the law help progressively develop how it regulates a domain of life. Appellate courts may confine themselves to checking the acceptability of the explanation. But first and foremost, the legal order cares about the context of discovery. It is concerned if jurisprudence discriminates against a group of plaintiffs or defendants, even if the ruling is justified in inoffensive terms. More fundamentally even, at least as a regulative idea, the law wants to do the parties justice. This requires that ambiguous cases are decided on their merits, not on extraneous grounds that interfere with judicial intuition. With this paper we can only alert the legal system to the possibility of judicial intuition being misled. All we can show is that, at least in the case we have tested, participants seem to be unaware of the counterintuitive effect that elaborating on counterarguments may have. It will take a whole line of research to uncover
the most important determinants of judicial intuition. Only then will one be in a position to decide whether, and if so which, countermeasures should be taken.

Every empirical project has limitations. The obvious limitation of our project is our inability to support our hypotheses. We would argue, though, that the hypotheses are not far-fetched, and that only through our attempt at validating the underlying conceptual claim we have found in which ways mental process and its reconstruction in the reasons given by the decision-maker may fall apart, in a run-of-the-mill torts case.

We have not tested professional judges, but students. Hence our participants had neither professional training nor professional experience. Yet tort cases are frequently decided by jury trial, i.e., by lay decision-makers. The processes we investigate, and even more so the basic finding of the paper, are very generic: those who make legal decisions are unaware of the mental process; the reasons they give are unrelated to the determinants of the decision-making process. At this level of generality, the fact that students are not representative for people on jury duty is unlikely to bias results.

One might be concerned that results are artefacts of differences in word length. Yet results stay essentially the same if we control for word length (Table S14). One might also be concerned that results are artefacts of the difficulty of words. Yet if we add the difficulty scores as controls that participants have given us, results also largely stay the same (see above, Table S12). One of the eye tracking devices did not properly discriminate between fixations. We have excluded these data from the analysis. Yet if we add them, results do not change in any relevant way (see above, Table S10).

Despite the limitations inherent in our method, we believe that, as our experiment demonstrates, eye tracking is a promising method for studying what seems so opaque for most observers: the ability of legal decision-makers to meaningfully decide about ill-defined cases.
References


### Appendix

<table>
<thead>
<tr>
<th></th>
<th>number of fixations</th>
<th>fixation duration</th>
</tr>
</thead>
<tbody>
<tr>
<td>causation</td>
<td>.825** (3.17)</td>
<td>1.041*** (3.54)</td>
</tr>
<tr>
<td>negligence</td>
<td>1.070** (2.82)</td>
<td>1.001* (2.36)</td>
</tr>
<tr>
<td>compensation</td>
<td>.390 (1.57)</td>
<td>.559** (2.97)</td>
</tr>
<tr>
<td>deterrence</td>
<td>.509 (1.42)</td>
<td>.406 (1.26)</td>
</tr>
<tr>
<td>contributory</td>
<td>.755*** (4.16)</td>
<td>.823*** (5.25)</td>
</tr>
<tr>
<td>excuse</td>
<td>-.008 (-.02)</td>
<td>-.144 (-.38)</td>
</tr>
<tr>
<td>excessive</td>
<td>.543* (2.17)</td>
<td>.534** (2.85)</td>
</tr>
<tr>
<td>chilling</td>
<td>.728*** (4.02)</td>
<td>.571*** (3.49)</td>
</tr>
<tr>
<td>N</td>
<td>115</td>
<td>115</td>
</tr>
<tr>
<td>RMSE</td>
<td>.349</td>
<td>.346</td>
</tr>
<tr>
<td>adj.R²</td>
<td>.752</td>
<td>.756</td>
</tr>
</tbody>
</table>

**Table 4**  
Predictive Power of Eye Gaze Even When Ignoring Treatment

OLS  
causation, negligence, compensation, deterrence, contributory, excuse, excessive, chilling:  
shorthands for items on the screen of Figure 1  
all items represented with proportion, per participant, of fixations on all 8 items / duration of these fixations  
models without constants, as proportions add up to 1 by design  
one (aggregate) datapoint per participant  
standard errors in parenthesis  
*** p < .001, ** p < .01, * p < .05, + p < .1

---

25
Stimulus Materials

Figure 5
Sequence of Tasks in Role of Judges

Figure 6
Sequence of Tasks in Role of Counsel
Vignette Judges

A earns 3,000 €/month. He does not own any property that could legally be seized by a creditor. A drives in his car and passes a village. The speed limit is 30 km/h. A drives at 55 km/h. A cat jumps on the street. A breaks to save the cat, and loses control of his car. The car runs into an above-ground rain water tank owned by B. The car runs into an above-ground fuel tank owned by B. After the tank has been installed, construction law has changed. Today, putting the tank above ground would no longer be allowed. An expert testifies that A would not have lost control had she respected the speed limit. A’s insurance does not cover damage to property on ground outside the street.
Repairing the tank costs 1,500 €. There is a risk of ground water contamination. This is why the police orders B to dredge the ground, at a cost of 43,500 €. B sues and claims 1,500 € 45,000 € in damages.

The plaintiff requests full compensation. The defendant argues that at most partial compensation (500 €, 15,000 €) can be justified.

Sketch the reasons for your decision.

Vignette Counsel

A earns 3,000 €/month. He does not own any property that could legally be seized by a creditor. A drives in his car and passes a village. The speed limit is 30 km/h. A drives at 55 km/h. A cat jumps on the street. A breaks to save the cat, and loses control of his car. The car runs into an above-ground fuel tank owned by B. After the tank has been installed, construction law has changed. Today, putting the tank above ground would no longer be allowed. An expert testifies that A would not have lost control had she respected the speed limit. A’s insurance does not cover damage to property on ground outside the street.

Repairing the tank costs 1,500 €. There is a risk of ground water contamination. This is why the police orders B to dredge the ground, at a cost of 43,500 €. B sues and claims 45,000 € in damages.

Write your plea.
[To the counsel of the plaintiff] Argue for full compensation (45,000 €).
[To the counsel of the defendant] Argue for reduced compensation (15,000 €).