Moral Thin-Slicing

Julian De Freitas
Harvard Business School

Alon Hafri
Johns Hopkins University

Working Paper 23-002
MORAL THIN-SLICING

AUTHORS ANONYMIZED

DATE SUBMITTED: November 1, 2022
ABSTRACT (196 WORDS [MAX: 200 WORDS])

Given limits on time and attention, people increasingly make moral evaluations in a few seconds or less, yet it is unknown whether such snap judgments are accurate or not. On one hand, the literature suggests that people form fast moral impressions once they already know what has transpired (i.e., who did what to whom, and whether there was harm involved), but how long does it take for them to extract and integrate these ‘moral atoms’ from a visual scene in the first place to decide who is morally wrong? Using controlled stimuli, we find that people are capable of ‘moral thin-slicing’: they reliably identify moral transgressions from visual scenes presented in the blink of an eye (< 100 ms). Across four studies, we show that this remarkable ability arises because observers independently and rapidly extract the atoms of moral judgment — event roles (who acted on whom) and harm level (harmful or unharmful). In sum, despite the rapid rate at which people view provocative moral transgressions online, as when consuming viral videos on social media or negative news about companies’ actions toward customers, their snap moral judgments about visual events can be surprisingly accurate.

Keywords: Moral judgment; thin slices; internet; event structure; visual relations; social media
On April 9th of 2017, a Vietnamese-American named Dr. David Dao Duy Anh was selected to be involuntarily deplaned from United Airlines Express Flight 3411. Dao, a pulmonologist, refused to surrender his seat because he had patients to see the next day. Given his resistance, security officers forcibly removed him from the plane and in the process struck Dao’s face on an armrest, apparently knocking him unconscious. Officers dragged Dao’s limp body through the row of onlooking passengers.

Some of those passengers used their mobile cameras to film the incident, upload it online, and distribute it via social networks, where it spread virally. Outraged consumers aggregated United’s other transgressions under the #BoycottUnited hashtag on Twitter (Zorthian, 2017), and major news sources scrutinized and criticized CEO Oscar Munoz’ accusation that Dao was “disruptive” and “belligerent” (McCann, 2017). Even U.S. President Donald Trump called the whole incident “horrible” (Wojcik, 2017). In little time, United’s stock dropped precipitously (Garber, 2017) and CEO Munoz publicly apologized, characterizing Dao as faultless. United and Dao settled ‘amicably’ (BBC, 2017).

Examples like this illustrate that, while it is rare these days for people to directly witness moral transgressions, such transgressions are more widely accessible today than ever before (Pinker, 2012). They can be captured, uploaded online, and distributed via social networks extending from the present into the future; aggregated in memes on social media and video-sharing sites (as in hashtags like #instant karma, #random acts of kindness, #dashcam scam, and others); and covered in the latest news online. Moral transgressions are also highlighted in emotionally evocative advertisements and political campaigns, and prevalent in simulated virtual environment games such as Second Life and Grand Theft Auto.

---

1 https://www.youtube.com/watch?v=VrDWY6C1178
The United Airlines example also illustrates that although these moral transgressions are witnessed online, they are consequential, leading to an increase in moral emotions such as outrage and a loss in happiness (Crockett, 2017; Hofmann, Wisneski, Brandt, & Skitka, 2014; Lindenmeier, Schleer, & Pricl, 2012), and punitive action against others such as snowballing blame on social media or visibly boycotting a brand (Klein, Smith, & John, 2004). These reactions require just a few keypresses from the comfort of a person’s home, yet they can have lasting effects like ruining a reputation, catalyzing a feud, or devaluing a firm. A single video can even start a social movement, as when the viral video of a police officer killing a Black US citizen named George Floyd catalyzed global protests.\(^2\)

Yet, given limits on time and attention, people increasingly make moral evaluations based on what they see in tens of milliseconds to around four seconds on any given webpage, news, or social media content (Decker et al., 2015; Maslowska, Ohme, & Segijn, 2021; Vraga, Bode, & Troller-Renfree, 2016; Weinreich, Obendorf, Herder, & Mayer, 2008). Large social media firms like Twitter and Facebook have even introduced ‘read before you share’ prompts (Ghaffary, 2021; Vincent, 2020) to slow readers down and reduce misinformation on their sites which might arise from such brief viewing behavior (Lazer et al., 2018).

Such interventions assume that people make less accurate moral inferences about content viewed for brief versus longer periods, yet to our knowledge no previous work has investigated whether this is true. The most related work has investigated the ability to make rapid judgments about non-social content likes ads or single individuals. These studies find that people actually require just 100 ms to accurately judge whether presented content is an ad or an editorial (Pieters & Wedel, 2012) or whether an individual face is attractive, likeable, trustworthy, competent, or

\(^2\) [https://en.wikipedia.org/wiki/List_of_George_Floyd_protests_outside_the_United_States](https://en.wikipedia.org/wiki/List_of_George_Floyd_protests_outside_the_United_States)
aggressive (Willis & Todorov, 2006). This ability to make judgments based on brief periods, or ‘thin slices’, of information is referred to as ‘thin slicing’, and suggests that people are more accurate at making speeded judgments than previously thought (Peracchio & Luna, 2006). Here we ask whether people are capable of ‘moral thin-slicing’, in which they require just a blink of an eye to tell whether a social interaction is harmful and who is the transgressor.

**Thin Slicing and The Computational Challenge of Moral Judgment**

Although people can make sub-second thin slice judgments in some contexts, they might not be able to do so for *moral* judgments, given that they are computationally complex: First, to make a moral judgment often one needs to know whether it is harmful (Schein & Gray, 2018) and who in the interaction is the doer (aka ‘Agent’) or recipient (aka ‘Patient’) (Hafri, Papafragou, & Trueswell, 2013). Generally, a person is morally wrong if they are involved in a harmful interaction and are the agent of that interaction. There are also other factors that affect moral judgment, such as the actor’s intentional states (Allison, Puce, & McCarthy, 2000; Dennett, 1989), causal factors leading to the event (De Freitas, DeScioli, Nemirow, Massenkoff, & Pinker, 2017; Tsiros, Mittal, & Ross Jr, 2004; Weiner, 2000), and whether the agent had good reasons for their actions (Malle, Guglielmo, & Monroe, 2014).

Second, moral judgment poses a formidable challenge for visual perception, because moral judgments are not just about the actions of a single individual, but about physical and social relations taking place between individuals. Thus, moral judgments of visually depicted social interactions involve integrating inferences about individuals into a whole (Hafri & Firestone, 2021). Furthermore, there is wide variation in what harm can look like and in the various possible configurations of individuals relative to each other.
Third, moral judgment is complex because it involves integrating information from multiple cognitive systems spanning vision (which is fast and automatic) and cognition (which is relatively slow and deliberative) (Kahneman, 2011). The visual system detects whether an interaction is social or non-social (Isik, Koldewyn, Beeler, & Kanwisher, 2017; Papeo, 2020; Papeo & Abassi, 2019) and determines whether each actor is the Agent or Patient of the social interaction by processing configural shape features like whether one person is leaning forward more than other person and has his arms outstretched (Hafri et al., 2013). In contrast, since there are no literal features in an image that trivially indicate moral wrongness, cognitive brain areas beyond the visual system, such as the ventromedial prefrontal cortex, might be required for integrating together abstract information to make moral judgments (J. Greene & Haidt, 2002).

**Moral ‘Atoms’**

Given these computational challenges, and the practical relevance of judging moral transgressors online, how quickly can people do this? Although moral transgressions can be witnessed in various formats, such as in the form of written gossip or video, we focus on the ubiquitous case of visually presented transgressions of one actor engaging with another, as in images shared on news, social media, or video-sharing sites (Figure 1a). Here we investigate whether observers can make a moral judgment about a briefly viewed visual scene by rapidly extracting and assembling together the ‘atoms’ of moral judgment: event role information (who acted on whom) and harm level (whether an action caused harm or did not).

Our work is different from the ‘social intuitionist’ model of moral judgment, which says that people’s moral judgments are often informed by a fast, intuitive assessment of a situation (Haidt, Bjorklund, & Murphy, 2000). For example, people’s moral judgments of whether it is
okay for two siblings to engage in sexual intercourse is driven more by their feelings of disgust than by reasoned thinking about whether anyone was harmed (Haidt, Koller, & Dias, 1993). That work and the current study have in common the question of whether moral judgments can be fast. However, the aspects of moral judgment under investigation in that work and the current study are quite different. While the social intuitionist model convincingly shows that a first-pass emotional reaction (disgust) informs moral judgment, this reaction only has its effect after people have already read and understood what has transpired, i.e., they have learned who acted on whom and what it was that they did. In contrast, here we study the speed with which people assemble these ‘atoms’ of moral judgment—event roles and harm level—in the first place. Thus, what the current work asks is whether, at least when presented with a visual scene, the (non-moral) atoms of moral judgment are extracted so quickly that they afford a rapid moral judgment.

But how rapid? One possibility is that people must view visual scenes for at least several seconds to accurately judge a moral transgressor. This would show that moral judgment is a relatively lengthy, deliberative judgment, akin to predicting a negotiation outcome. Indeed, even many studies under the banner of ‘thin slicing’ have actually asked people to evaluate content presented not for mere milliseconds but for several seconds or even minutes (Ambady & Rosenthal, 1992), such as predicting a negotiation outcome based on the first five minutes of the interaction (Curhan & Pentland, 2007), or trustworthiness of a salesperson based on a 30-second clip (Ambady, Krabbenhofst, & Hogan, 2006; Hall, Ahearne, & Sujan, 2015; Main, Dahl, & Darke, 2007).

A second, more exciting, possibility is that people need to observe a social interaction for less than a blink of an eye (< 100 ms), which would show that some moral judgments rely on
relatively rapid, automatic processes, more akin to judging the trustworthiness of a face or
distinguishing an ad from an editorial (Pieters & Wedel, 2012; Willis & Todorov, 2006).

Providing some tentative evidence toward the second possibility, previous work finds that
people rapidly detect the ‘inputs’ on which moral judgment depends. One study found that
people presented with naturalistic photographs of social interactions needed just 37 ms to
accurately recognize the social interaction’s category (e.g., ‘kicking’ or ‘punching’) and roles of
the actors engaged in it (i.e., Agent or Patient) (Hafri et al., 2013). Another study found that
visually processed information about causation affected peoples’ moral judgments even when
people believed it did not. De Freitas & Alvarez (2018) showed participants events in which it
was ambiguous whether (a) one car hit into another car, which in turn hit into a pedestrian, or (b)
the second car accelerated on its own into the pedestrian. By including simple, task-irrelevant
peripheral events that participants did not believe affected their judgments, the display was made
to look more causal, which in turn increased moral blame of the driver of the first car.

As for information about harm, it is currently unknown whether people extract harm
information from a brief glance. With that said, previous work has found that people may rapidly
extract the “gist” of an event, i.e., its basic-level category (e.g., ‘kicking’ or ‘tickling’; Hafri et
al., 2013), suggesting that they should also be able to rapidly categorize events as harmful or not
harmful.

Altogether, however, these findings suggest that the visual system quickly provides the
high-level inputs on which moral judgment depends—such as social interaction category, event
role, and causality—enabling people to make snap moral judgments. Thus, although moral
judgment is typically considered a slower, ‘System 2’ process (Grunert, 1996; Kahneman, 2011),
in the case of visually presented transgressions, it may be very quick in practice because it leverages inputs that are rapidly extracted by ‘System 1’ perceptual processes.

**Assembling Moral ‘Atoms’**

Crucially, even if information about event role and harm is available to an observer, it is currently unknown whether observers can *integrate* these disparate types of morally relevant event information to make a moral judgment from a brief glance. Indeed, it is entirely possible that information about these moral ‘atoms’, despite being independently extracted at rapid speeds, are cognitively segregated and are only integrated when observers reflect about an event effortfully and deliberatively. This situation could arise, for example, if the two types of information are output in different representational formats (e.g., imagistic vs. abstract; Marr, 1982), or are processed by two different visual streams (D. Milner & Goodale, 2006; Ungerleider & Mishkin, 1982).

Information segregation is evidence in both visual and cognitive processing, as in the “what”/“where” (or “what”/“how”) division of processing in visual processing (D. Milner & Goodale, 2006; Ungerleider & Mishkin, 1982). Indeed, patients with damage to the ventral stream show selective deficits in representing object information but have intact object localization or action information, while those with damage to the dorsal stream show the reverse effects (D. Milner & Goodale, 2006; Ungerleider & Mishkin, 1982). Even neurotypical individuals exhibit signs of this information segregation, as when they experience a perceptual size illusion while still accurately adjusting their fingers to the correct size of the misperceived object when reaching for it (Rossetti, 1998). Thus, information represented in a segregated manner can fail to be integrated successfully toward a common behavioral goal.
Another example of segregated information not being readily integrated comes from spatial tasks which find that cognitive load can prevent proper integration of disparate information. In these tasks, participants are disoriented in unfamiliar environments and must reorient to find rewards. Crucially, the environments are designed such that geometric information (i.e., the overall ‘shape’ of the environment) is insufficient to disambiguate the correct location; non-geometric information is also required (e.g., the color or texture present at certain locations). Young children and non-human animals fail to use the non-geometric cues, and instead rely primarily on local geometry to reorient: they search not only in the correct location but also in rotationally equivalent locations (Cheng, 1986; Hermer & Spelke, 1994; Julian, Keinath, Muzzio, & Epstein, 2015). Strikingly, adults under linguistic interference (i.e., verbal shadowing) show similar error patterns, despite being able to detect and remember such information (Hermer-Vazquez, Spelke, & Katsnelson, 1999).

Collectively, these examples demonstrate situations in which disparate kinds of visual information fail to be integrated successfully towards a common behavioral goal, i.e., recognizing or grabbing an object, or reorienting in an environment. Relating to the current investigation, observers of visual events may rapidly represent role and harm information without also rapidly integrating them to make moral judgments, perhaps because the visual information remains segregated unless deliberated upon.

‘Moral Thin-Slicing’?

In contrast to the possibility that morally relevant information processed from a brief glance at a visual scene remains cognitively segregated, here we hypothesize that observers are capable of ‘moral thin-slicing’: they can accurately judge who is the moral transgressor in
images presented for less than 100 ms. We further propose that the speed of moral thin-slicing depends on a ‘perceptual bottleneck’: how quickly people can extract the inputs (the atoms) on which moral judgment depends, i.e., event role and whether the interaction was harmful or not. By the same token, causally increasing how challenging it is to extract event role and harm information leads to decrements in the accuracy of moral judgments.

OVERVIEW OF STUDIES

We investigate the possibility of ‘moral thin-slicing’—accurate moral judgments of images presented for < 100 ms. Our studies employed a controlled image set (Figure 1b) to determine whether people are truly capable of moral thin-slicing without other confounding factors, such as scene context, social identities (e.g., bouncer vs. patron), or viewpoint. Given images of actors engaged in either harmful or unharmful social interactions, participants were tasked with determining who acted on whom (‘Role’), whether harm was inflicted (‘Harm’), and how morally wrong each actor was (‘Moral Wrongness’). Study 1 (‘Moral Thin-Slicing’) established the basic moral thin-slicing effect as well as a victim-blaming effect under speeded presentation, Study 2 (‘Temporal Evolution of a Moral Judgment’) traced the evolution of a moral judgment from 17 ms to 1500 ms, and Study 3 (‘Causal Manipulation of Role’) and Study 4 (‘Causal Manipulate of Harm’) causally intervened on the psychological processes.
Figure 1. Example images of moral transgressions, in the real-world and our controlled experiments. (a) Participants view images of many kinds of low- and high-harm interactions every day in social media, news, video-games, and other simulated environments. (b) Controlled image set in which identical twins perform harmful and unharmful interactions, with side and agent fully counterbalanced. Sources for images in (a), from left-to-right by row:


STUDY 1: MORAL THIN-SLICING

Using a carefully normed image dataset (see Study S1 in Supplemental File), Study 1 tested whether participants accurately detect moral information when presented in the blink of an eye.
Method

Participants. We collected data from 134 participants from Amazon’s Mechanical Turk (Mturk), which was based on sample sizes of previous online studies with similar display characteristics (Hafri, Wadhwa, & Bonner, 2022). The study link specified that participants should have normal or corrected-to-normal visual acuity. Participants in all studies were allowed to participate if they had an Mturk approval rating above 95%, had participated in at least 100 studies previously, were located in the U.S.A, and had not taken part in any previous study in this project. These criteria were designed to select for motivated online study participants with normal or corrected-to-normal vision naïve to the purposes of the studies. We additionally included several attention checks and monitored the timing of stimuli to ensure reliability of data. Thirty-one participants were excluded from analysis based on exclusion criteria outlined below.

We took no special measures beyond these for recruiting participants, and thus the sample was as diverse and inclusive as the population of online participants that met the above recruitment criteria at the time of data collection, which is generally more diverse than the typical college undergraduate (Mason & Suri, 2012; Peer, Brandimarte, Samat, & Acquisti, 2017). Beyond the above criteria, we have no reason to believe that the results reported here depend on characteristics of the participants not considered above.

Stimuli. We employed 108 images of identical-twin actors engaged in 27 common social interactions varying in the degree of harm: bandaging, brushing, calling after, feeding, cover, poking, lifting, tickling, look at, face-painting, hugging, kissing, dressing, tapping, filming,
strangling, shooting, kicking, punching, slapping, tripping, scratching, stabbing, biting, scaring, listening to, pulling.

For each social interaction category, we counterbalanced the spatial location of the agent as well as the colors of their shirts (blue or orange-red, so that they would appear distinct even to color-blind individuals). Otherwise, the actors were identical twins (age 29) with identical haircuts and clothes. The actors were photographed against a plain light-blue background, then the background was post-processed to a uniform level of brightness. In images where an instrument was held (e.g., a knife, in the stabbing category), both the actor (aka ‘agent’) and recipient (aka ‘patient’) held duplicates of the instrument. For more information see Hafri, Papafragou & Trueswell (2013).

Procedure. After consenting, participants were redirected to a web server where platform-independent stimulus presentation and data collection were completed by custom software run in their web browsers, written using a combination of html, CSS, and jQuery.

Each participant saw all the images, but the task they completed in response to these images varied between-subjects, consisting of the following questions: (i) Color: “Was the person on the LEFT (RIGHT) wearing a red shirt?”, (ii) Role: Was the person on the LEFT (RIGHT) acting on the other person?, (iii) Harm: Was there harm being inflicted?, (iii) Moral Wrongness: Was the person on the LEFT (RIGHT) doing something morally wrong? For the moral wrongness task, participants were asked about the agent for half of the images, and the patient for the other half. For all tasks, the answer options were always “yes” or “no”. Which side (left or right) participants were asked about was counterbalanced within-subjects for the color, role, and moral wrongness tasks. The color task was intended as a non-social baseline for
which we expected participants to perform well. We aimed for N = 30 in each task condition. For the moral wrongness task, we expected agents and patients to elicit categorically distinct moral judgments (Gray, Waytz, & Young, 2012), so for this task we collected approximately twice as many participants as in the other conditions, in order to maintain the same statistical power for each agent type.

Because the display loaded within participants’ own web browsers, viewing distance and screen resolutions could vary dramatically, so we report dimensions of the stimuli using pixel (px) values and positions of the stimuli as pixel values relative to the left and top borders of a gray (red [R]: 221, green [G]: 221, blue [B]: 221) task window (800 px × 554 px), within which the images were presented.

The gray task window was always present on the screen. At the beginning of a trial, the word “Ready?” (font-size: 14pt) appeared for 400 ms in the middle of the task window, followed by a fixation cross (font-size: 14 pt) that appeared for 100 ms in the same location, followed by the image (590px × 443px; left: 105px, top: 3px) which was shown for 33 ms before disappearing. The task question (615 px x 10 px; left: 100 px, top: 446 px; font-size: 15 pt) stayed on-screen until participants pressed either the “y” or “n” keys to answer yes or no (figures 2a-b), else the screen timed out after one minute.

Exclusions. Three social interaction categories — scaring, listening to, and pulling — were excluded based on the unspeeded norming study described in Study S1 in the Supplementary Materials. Specifically, these categories were excluded because they had low response agreement on at least one of the four tasks (i.e., more than 2.5 SDs below each task’s mean agreement). The remaining 24 categories had high average agreement rates across all tasks.
We note that results of this study are qualitatively the same whether or not these three categories were excluded from analyses; in other words, results were statistically significant and in the same direction in both cases.

Participants were excluded for failing the comprehension/attention checks, or if they indicated that they completed a similar task (i.e., they responded “yes” to the post-experiment question: “Have you ever completed a HIT containing a similar scenario, perhaps involving the same sorts of questions?”). We also excluded participants who lost >=15% of their trials due to trial-based exclusion criteria. A trial was excluded from the study if response time was < 150 ms (suggesting that the observer was holding down a key) or > 1 minute (suggesting inattention). In addition, we recorded presentation durations for each trial using standard JavaScript timing functions and excluded trials that did not meet one of the following timing criteria: (i) the image was presented for < 25 ms or > 40 ms (as measured by the browser), rather than for the intended duration of 33 ms, or (ii) the browser refresh rate was measured at < 30 fps or > 500 fps (suggesting display timing issues). Note that although we excluded data for the above reasons, the results reported below do not depend on these trial or participant exclusions; in other words, all effects remained qualitatively the same (were statistically significant in the same direction) regardless of whether data from these excluded trials and participants were included or not — and the same is true for every experiment reported in this paper.

Transparency and openness. We report how we determined our sample size, all data exclusions (if any), all manipulations, and all measures in the study. All data and analysis code have been made publicly available on Github and can be accessed at [persistent URL or DOI].
Data were analyzed using R, version 4.0.2 (R Core Team, 2020). Study design and analyses were not pre-registered. This statement applies to all studies reported in this manuscript.

**FIGURE 2**

EXPERIMENTAL PARADIGM (A) AND PROBE CONDITIONS (B) IN STUDY 1

NOTE.— (a) Participants viewed an image for 33 ms (unmasked) and were then presented with one of four possible probes. (In the pre-experiment norming study S1, the image and probe appeared at the same time, and remained on screen until response.) (b) Four probe conditions, shown between-subjects. Participants were asked about the color, role, or moral wrongness of one of the two people in the image (the one on the left or right, asked within-subjects), or about whether the action was harmful or unharmful. In the norming Study S1, the social interaction in (a) stayed on screen until the participant responded. The same paradigm with different stimuli was used for Studies 3–4.

Results

To measure whether participants accurately extracted morally relevant information about the images, we calculated $d$-primes ($z$-transformed hit rate – false alarm rate) for each participant and then tested significance across participants relative to zero (chance), doing so for each task. Despite the speeded presentation, participants accurately extracted color, role, and harm information (Color: $t(22) = 9.57, p < .001, d = 2.00$; Role: $t(19) = 9.47, p < .001, d = 2.13$; Harm: $t(20) = 8.92, p < .001, d = 1.95$). Furthermore, as hypothesized, they accurately made moral wrongness judgments ($t(40) = 10.83, p < .001, d = 1.69$), suggesting that they extracted both role and harm information in concert and successfully integrated them to make moral wrongness
judgments (Figure 3).

**FIGURE 3**

RESULTS FROM STUDY 1

![Graph showing mean d-primes across participants for each judgment condition, with error bars representing 95% confidence intervals. Horizontal lines above conditions reflect pairwise significance tests between each condition, corrected for six between-task comparisons using the Bonferroni-Holm method. ***p < .001; * p < .05; n.s. not significant, p's > .112.]

**NOTE.**— Participants were above-chance at extracting morally relevant information, including role, harm, and moral wrongness. Points are mean d-primes across participants for each judgment condition, and error bars are 95% confidence intervals. Horizontal lines above conditions reflect pairwise significance tests between each condition, corrected for six between-task comparisons using the Bonferroni-Holm method. ***p < .001; * p < .05; n.s. not significant, p's > .112.

Mean categorization performance on the tasks, from highest to lowest, was as follows: color, role, harm, and moral wrongness. We conducted two complementary statistical analyses: We first conducted independent-samples t-tests to test which conditions differed from one another. Below, we report uncorrected $p$ values ($p_{unc}$), and $p$ values corrected for six between-task comparisons using the Bonferroni-Holm method ($p_{cor}$). To complement these analyses, we also calculated two-sample Bayes Factor $t$ tests between each condition and the next (using the R package *BayesFactor* with the function *ttestBF* and the default medium prior of $\sqrt{2}/2$). We looked for evidence in favor of the null hypothesis of no difference (i.e., $BF_{01}$), which would be evinced by a $BF_{01} > 1$. 
Crucially, these tests revealed that Role differed significantly from Moral Wrongness ($t(30.46) = 3.19, p_{unc} = .003, p_{cor} = .013, BF_{01} = 0.03$), while Harm did not ($t(40.26) = 0.98, p_{unc} = .335, p_{cor} = .335, BF_{01} = 2.48$) (see also Figure 4). They also revealed that Color differed significantly from Harm ($t(34.45) = 3.94, p_{unc} < .001, p_{cor} = .002, BF_{01} < 0.01$) and Moral Wrongness ($t(29.30) = 4.83, p_{unc} < .001, p_{cor} < .001, BF_{01} < 0.01$), and that the other conditions did not differ significantly from one another: Color vs. Role: $t(39.24) = 2.01, p_{unc} = .052, p_{cor} = .112, BF_{01} = 0.74$; Role vs. Harm: $t(35.86) = 2.16, p_{unc} = .037, p_{cor} = .112, BF_{01} = 0.52$.

The fact that Role differed from Moral Wrongness while Harm did not suggests that identifying harm served as a perceptual ‘bottleneck’ on performance for the moral wrongness task involving these stimuli, in line with our hypothesizing. This interpretation is reinforced when we look at item-by-item accuracy, where we find that accuracy on the harm task, but not the role task, significantly predicts accuracy on the moral wrongness task (harm: $r(22) = 0.91, p < .001$; role: $r(22) = -0.06, p = .791$; Figure 4a). We also note that this does not simply mean that harm and moral wrongness are redundant, since making an accurate moral wrongness judgment relies on integrating harm information with role information, e.g., even if you know an interaction is harmful, in order to accurately say whether a given actor is morally wrong you also need to know whether they are the agent or patient of the interaction.

Next, we explored whether participants truly integrated role and harm information to make moral wrongness judgments, by conducting a repeated measures ANOVA, with role type (agent blame vs. patient blame) and harm level (harmful vs. unharmful event) as factors. If role and harm information both informed moral wrongness judgments, we should find an interaction of role type and harm level, such that agents in harmful interactions are rated as more morally wrong than agents in unharmful interactions.
Supporting this interpretation, we found significant main effects of role type (agent vs. patient; $F(1,40) = 90.59, p < .001, \eta^2_p = 0.87$), harm level (harmful vs. unharmful event; $F(1,40) = 100.2, p < .001, \eta^2_p = 0.80$) and a significant interaction between the two ($F(1,40) = 36.75, p < .001, \eta^2_p = 0.48$). In post-hoc tests, we found that agents were rated as significantly more morally wrong than patients (68.3% of the time for Agents, 20.2% for Patients, $t(40) = 11.69, p < .001, d = 1.83$; Figure 4b), and agents in harmful interactions were rated as more morally wrong than agents in unharmful interactions (78.2% of the time for Agents in harmful interactions, 37.6% for Agents in unharmful interactions, $t(40) = 11.58, p < .001, d = 1.81$; Figure 4b). Thus, despite the speeded presentation and the fact that participants were never explicitly asked to base their moral judgments on role or harm, moral wrongness judgments leveraged information about both.

We also separately examined moral wrongness effects for the patient. For unspeeded durations (reported in study S1 in the Supplementary Material), we found that patients were rated less morally wrong than agents for all social interaction categories. Yet we were curious whether this difference required more viewing time to emerge or was present even for speeded presentations. To test this question, we conducted a mixed-effects ANOVA with harm level (harmful vs. unharmful interactions, within-subject) and speed condition (speeded vs. unspeeded, between-subjects) as factors. We found no significant effect of speed condition (speeded vs. unspeeded; $F(1,102) = 3.12, p = .080, \eta^2_p = 0.06$), nor an interaction of speed condition and harm level ($F(1,102) = 1.40, p = .240, \eta^2_p = 0.01$), although we did find a main effect of harm level (harmful vs. unharmful event; $F(1,102) = 40.56, p < .001, \eta^2_p = 0.29$). Post-hoc analysis found that patients were attributed more harm for participating in a harmful event (27% of the time) than an unharmful one (8.6% of the time, $t(40) = 6.36, p < .001, d = 0.62$; Figure 4b), supporting our hypotheses. Indeed, for patient moral wrongness, the highest rating for an unharmful
interaction (*tap*, *M*=0.13) was still not as high as the lowest rating for a harmful interaction (*bite*, *M*=0.20). This effect is reminiscent of so-called victim blaming effects (Ryan, 1976), although to our knowledge this is the first time that the effect has been demonstrated for presentations of social interactions that are presented so briefly and that are so generic, suggesting that victim blaming could be a default bias in moral judgment.

Discussion

In line with our hypothesis, Study 1 found that people make accurate moral judgments about social interactions based on ‘thin slices’ of viewing time. They do this by extracting both role and harm information about the image and combining them to make a moral judgment. We also found that the speed at which they made moral judgments depended on how quickly they extracted harm information, in line with Hypothesis 2a.

Interestingly, we also found a victim blaming effect, whereby participants judged patients as more morally wrong when they were involved in harmful than unharmful interactions. Previous studies of victim blaming have asked participants to make moral judgments about richer social scenarios, e.g., as conveyed through moral vignettes, questionnaire data, thought experiments like the so-called trolley problem (De Freitas et al., 2017; Lerner & Miller, 1978; Niemi & Young, 2016), or unspeeded images or video (Bohner, 2001; Hafer & Begue, 2005). Here, we find an analogous effect whether observers view brief or long presentations of quite generic social interactions, suggesting that victim blaming may be rooted in a general, default way of categorizing social interactions. We return to this issue in the General Discussion.
NOTE.— (a) Pearson correlations of action-by-action accuracy on Agent blameworthiness with accuracy on role (left panel) and harm (right panel), at brief display. Accuracy on moral wrongness was highly correlated with accuracy on harm, but not role, suggesting that harm extraction was the “bottleneck”: correctly judging moral wrongness was contingent on correctly identifying whether an action was harmful. (b) Moral wrongness ratings of the Agent and Patient at brief and long displays (Study 1 and S1, respectively), for unharmful (left panel) and harmful (right panel) interactions. At brief displays, Agents were blamed more than Patients, and this difference was stronger for harmful relative to unharmful social interactions, suggesting that even with minimal perceptual information, Harm and Role information were integrated to perform moral wrongness judgments. Additionally, at both brief and long displays, Patients in harmful interactions were judged as more morally wrong than Patients in unharmful ones, suggestive of a general ‘victim blaming’ effect. Only the significance of this victim blaming difference is indicated in the plot (for other significant effects, see main text). Points are means across participants, and error bars are 95% confidence intervals. *** \( p < .001 \).

STUDY 2: TEMPORAL EVOLUTION OF A MORAL JUDGMENT

Just how quick is moral thin-slicing? What is the minimal amount of time a person needs to view a visual scene to make an accurate moral judgment, and what determines this speed?
Study 2 explored this question for each of the tasks, by testing performance at 15 different presentation durations ranging from 17 ms to 1500 ms. We planned to test how long an image must be visible before performance at a given speeded presentation is indistinguishable from performance at the unspeeded presentation.

Since harm and role information are the ‘atoms’ of a moral judgment, the speed of a moral judgment should depend on how quickly an observer can extract both role and harm information. Or put another way, role and harm extraction should serve as the ‘temporal bottleneck’ on speedy moral judgments, such that a moral judgment cannot be faster than the most slowly extracted information on which the judgment depends.

Method

We collected data from 2341 participants from Mturk and excluded 218 using the same criteria as in Study 1. The design was identical to Study 1, except that the images were presented for one of the following durations, between-subjects: 17, 33, 50, 67, 83, 100, 133, 150, 167, 200, 250, 500, 750, 1000, 1500.

Results

As in Study 1, we calculated $d$-prime values for each participant. We first verified whether we replicated Study 1, by testing for above-chance performance on each task at the 33 ms duration. Again, participants accurately made color, role, harm, and moral wrongness judgments at the 33 ms duration (Color: $t(26) = 20.24, p < .001, d = 3.90$; Role: $t(25) = 15.38, p$
Furthermore, the general order of task performance was similar to that of Study 1: color performance was greater than that for all other judgments, role was greater than harm and moral wrongness performance, and harm was greater than moral wrongness performance: Color vs. Role: $t(51.00) = 3.71$, $p_{unc} < .001$, $p_{cor} = .002$, $BF_{01} = 0.02$; Color vs. Harm: $t(50.54) = 6.57$, $p_{unc} < .001$, $p_{cor} < .001$, $BF_{01} < 0.01$; Color vs. Moral Wrongness: $t(55.87) = 9.52$, $p_{unc} < .001$, $p_{cor} < .001$, $BF_{01} < 0.01$; Role vs. Harm: $t(49.60) = 2.58$, $p_{unc} = .013$, $p_{cor} = .013$, $BF_{01} = 0.25$; Role vs. Moral Wrongness: $t(55.04) = 5.53$, $p_{unc} < .001$, $p_{cor} < .001$, $BF_{01} < 0.01$; Harm vs. Moral Wrongness: $t(65.63) = 3.25$, $p_{unc} = .002$, $p_{cor} = .004$, $BF_{01} = 0.08$.

Next, to test which speeded durations elicited the same level of performance as the unspeeded presentation, we calculated two-sample Bayes Factor $t$ tests between the unspeeded condition at each speeded duration condition. We looked for evidence in favor of the null hypothesis of no difference (i.e., $BF_{01}$’s > 1). Evidence for no difference at a given image viewing duration would indicate that the judgment performance at that viewing time reached the level of performance when viewing time was unconstrained. We expected that one of either role or harm extraction should require the same amount of viewing time as moral wrongness, indicating that it served as a perceptual bottleneck on the ability to make the moral wrongness judgment.

Participants detected role more rapidly than they did harm. Specifically, performance on the role task already reached the unspeeded level of performance when the images were presented for just ~67 ms, and by 150 ms, performance never dipped below the unspeeded level (Figure 5; see table S1 in Supplemental Material for statistical tests). In contrast, harm and moral
wrongness categorization only reached the unspeeded level of performance when images were presented for ~500 ms. The fact that performance for moral wrongness only reached unspeeded levels when harm did is in line with the perceptual bottleneck account—a moral judgment is only as fast as the slowest extracted piece of information on which it depends. Finally, while Color task performance was always higher than the other tasks at all speeds, it did not reach unspeeded levels until ~1000 ms, which we attribute to the near-ceiling performance of unspeeded participants on the color task. It is possible that since performance on color was so high for unspeeded presentation, any slight detriment from speeded presentation hurt color performance most.

**FIGURE 5**

RESULTS FROM STUDY 2

![Graph showing results from Study 2](image)

**NOTE.**— Images were displayed at durations ranging from 17 ms to 1500 ms. Unspeeded (unlimited time) results from Study 1 are shown for comparison. Points are mean d-primes across participants for each judgment condition. Horizontal lines above the plot indicate the durations at which performance on a given task at speeded durations showed no difference from the unspeeded condition (i.e. Bayes Factor > 1). By between 67 and 150 ms, participants performed as well on the Role task as they did under unspeeded presentation, while performance on the Harm and Moral Wrongness tasks did not reach unspeeded levels until at least 500 ms. Error bars are 95% confidence intervals.
Conclusion

This study tracked the temporal evolution of a moral judgment, showing that, in line with our predictions, the ability to accurately judge moral wrongness depends on the speed at which participants accurately extract the information on which moral judgment depends. Although participants already extracted role information as accurately as unspeeded levels after 67 ms of image presentation, they only did the same for harm information after 500 ms, at which point they could also do the same for moral wrongness. We note that although the order in which each ‘atom’ of the moral judgment emerged for these images (i.e., role first, then harm) might not generalize to other kinds of images (e.g., ones with more varied scene contexts or viewpoints), this is not as crucial as the fact that (i) different inputs to moral judgment may be extracted at different relative speeds, and (ii) extracting these inputs serves as the temporal bottleneck on accurately making moral judgments under speeded presentation.

STUDY 3: CAUSAL MANIPULATION OF ROLE

If participants are truly using role information to make moral judgments under speeded presentation, then causally manipulating how easy it is to extract role from the images should affect the accuracy of moral judgments. To this end, Study 3 showed participants a new set of manipulated images in which the patient of the social interaction leans forward with his limbs outstretched (making him a ‘non-prototypical’ patient). Typically in social interactions, it is the agent who has such postural characteristics (Hafri et al., 2013), so our goal here was to make it more challenging for an observer to distinguish the agent from the patient than in Studies 1–2, in
which the actor postures were more prototypical.

Method

We collected data from 301 participants from Mturk and excluded 37 using the same criteria as Study 2. The design was identical to Studies 1 and 2, except for the images used; i.e., Study 3 included both Speeded and Unspeeded tasks (between-subjects), with the analyses of interest being for the Speeded task. The agents in these images were staged identically to those in Studies 1–2, but the patient now had similar head orientation, body orientation, extremities, and body lean to the agent, making it more challenging for an observer to distinguish the agent from the patient under speed. Examples can be viewed in Figure 6a. For more image details see Hafri, Papafragou, & Trueswell (2013).

Since the aim of this study was to determine whether causally manipulating the difficulty of extracting role information affects judgments of moral wrongness under speed, we needed to ensure that participants could still tell who the agent was when given ample viewing time, i.e., we needed to ensure that our manipulation did not make it seem as though the patient was, in fact, the agent. To this end, we excluded from the analysis any social interactions for which unspeeded performance on non-prototypical patient images (Study 3) deviated significantly (> 3.0 SD) from the distribution of accuracy differences between Study 3 and Study S1 (original unspeeded norming study, available in the Supplementary Material). This way, the only causally manipulated images that are included are those that are categorized similarly as the unmanipulated images from Study S1 on all tasks (Color, Role, Harm and Moral Wrongness), when the participant has unlimited time to categorize them. (However, it is worthwhile to note
that these exclusions, if anything, should *hurt* our ability to detect differences in performance based on this role manipulation, as it requires that role information which is discernible to a similar degree in the two image categories under ample viewing time is less discernible under speeded viewing.)

This exclusion criterion led to the exclusion of two social interactions: *look at* and *tap*. As a sanity check, we confirmed that after this exclusion there were no significant differences in unspeeded performance between Study S1 and the current study on any task (Color: BF$_{01} = 1.78$; Role: BF$_{01} = 3.02$; Harm: BF$_{01} = 2.58$; Moral Wrongness: BF$_{01} = 4.63$). With that said, we also note that none of the results reported here depended on these exclusions, i.e., all effects reported below remain statistically reliable and in the same direction even without excluding these social interactions.

Results

To test whether manipulating role information causally affected moral wrongness performance, we planned to compare performance on the speeded task on Study 1 (prototypical patient) and the current study (non-prototypical patient). We predicted that with brief viewing time, the role manipulation would impact role performance and thereby moral wrongness performance (since it should depend on role extraction), while the other tasks should remain relatively unaffected.

Compared to participants shown the prototypical (Study 1) patient images, participants shown the non-prototypical images performed significantly worse at the role task ($t(40.30) = 2.11, p = .041, d = 0.62, BF_{01} = 0.58$) and moral wrongness task ($t(65.34) = 2.35, p = .022, d = 0.62, BF_{01} = 0.58$)
0.53, BF01 = 0.50), in line with our hypotheses. At the same time, participants did not perform significantly worse at the Color (t(30.41) = 0.34, p = .736, d = –0.10, BF01 = 3.38) or Harm tasks (t(33.84) = 0.89, p = .377, d = 0.26, BF01 = 2.43), suggesting that the harm and role tasks relied on non-overlapping visual features (Figure 6b).

**FIGURE 6**
RESULTS FROM STUDY 3

(a) In Study 3, the Patient’s posture was manipulated such that he was a non-prototypical Patient, possessing Agent-like postural features (Hafri et al., 2013). This manipulation was expected to disrupt Role extraction, and to lead to concomitant disruptions to Moral Wrongness extraction. (b) Study 3 results are plotted alongside data from Study 1. Points are mean d-primes across participants for each judgment condition and error bars are 95% confidence intervals. Horizontal lines above conditions reflect pairwise significance tests between each study. Relative to Study 1, Study 3 showed significant reductions in Role and Moral Wrongness extraction. * p < .05; n.s. not significant, p’s > .459.

Conclusion
In line with our predictions, causally manipulating the accessibility of role information selectively impaired performance on the wrongness task, confirming that role information is a direct input to moral judgment and can affect moral judgment independent of harm information.

**STUDY 4: CAUSAL MANIPULATION OF HARM**

Study 4 aimed to causally manipulate harm information. If participants truly use harm information to make moral judgments under speeded presentation, then, akin to Study 3, causally manipulating how easy it is to extract harm from the images should affect the accuracy of moral judgments. To decide on a manipulation, we were informed by the results of Study 2, which suggested that harm information takes longer (~500 ms) to emerge than other information—presumably because processing fine-grained details in these images takes more time. Thus, we reasoned that making the fine-grained details of the images harder to detect by darkening them would impair harm extraction. Although darkening an image makes all information harder to detect, we predicted that it would lead to less of a decrement in role extraction as compared to harm extraction, since role extraction relies on more global configural features rather than fine-grained details that are particularly affected by image darkening. To this end, Study 4 showed participants a separate set of manipulated images that were darkened versions of the original images from Study 1.

Method
We collected data from 292 participants and excluded 27 using the same criteria as in Study 1. Participants were recruited from the online platform Prolific (for a discussion of this and other online subject pools, see Peer et al., 2017). The design was identical to Studies 1–3, except that the luminance level of the original images was reduced by 80% (see Figure 7a for examples). This study included both Speeded and Unspeeded tasks (between-subjects), with the analyses of interest being for the Speeded task.

As in Study 3, we excluded from the analysis any social interactions for which unspeeded performance on the darkened images (Study 4) deviated significantly (> 3.0 SD) from the distribution of accuracy differences between Study 4 and Study 1 (the original images). This led to the exclusion of three social interactions: strangle, bite, and poke. Since we also planned to compare Study 4 to Study 3, we additionally excluded the social interactions that were excluded in the previous Study 3 analyses: look at and tap. After these action exclusions, there were no significant differences in unspeeded performance on any condition between Studies 4 and 1 (Color: BF$_{01}$ = 3.81; Role: BF$_{01}$ = 3.33; Harm: BF$_{01}$ = 2.87; Moral Wrongness: BF$_{01}$ = 4.66). And even with this larger set of action exclusions, there were still no differences in unspeeded performance between Studies 3 and 1 (Color: BF$_{01}$ = 1.34; Role: BF$_{01}$ = 2.90; Harm: BF$_{01}$ = 1.89; Moral Wrongness: BF$_{01}$ = 4.44). This exclusion procedure left 19 social interactions total in our analysis. However, as for the analyses of Study 3 reported above, we note that none of the results reported here depended on these exclusions, i.e., all effects reported below remain statistically reliable and in the same direction, even without excluding these social interactions.

Results
We planned to compare performance for the darkened images (the current study) to that for both the original images (Study 1) and non-prototypical patient images (Study 3). For each task, we ran a one-way ANOVA on $d$-prime values of the Speeded tasks only, followed by post-hoc two-sample $t$-tests separately comparing data from the current study with Study 1 and Study 3 (and we corrected for two multiple comparisons for each of the two planned cross-study comparisons, using the Bonferroni-Holm method).

Results can be seen in Figure 7b. For all tasks, we found either significant or near-significant differences among the three studies: Color ($F(2,70) = 50.00, p < .001, \eta_p^2 = 0.59$), Role ($F(2,67) = 2.82, p = .067, \eta_p^2 = 0.08$), Harm ($F(2,71) = 10.67, p < .001, \eta_p^2 = 0.23$), and Moral Wrongness ($F(2,117) = 13.21, p < .001, \eta_p^2 = 0.18$).

Firstly, post-hoc comparisons revealed that the current study showed a reduction in performance for the Color control task relative to the other two studies under speeded presentation (vs. Study 1: $t(33.12) = 7.40, p_{unc} < .001, p_{cor} < .001, d = 2.20, BF_{01} < 0.01$; vs. Study 3: $t(42.37) = 12.86, p_{unc} < .001, p < .001, d = 3.69, BF_{01} < 0.01$). This result replicates well-established findings of poor color vision under dim illumination (Pokorny, Lutze, Cao, & Zele, 2006), and so does not necessarily indicate a problem with the key aim of the darkening manipulation, especially since the moral wrongness task does not ask about color (participants are asked to morally judge the actor on either the left or right, not to judge an actor based on whether they are wearing a red or blue shirt).

Secondly, post-hoc comparisons for role, harm, and moral wrongness revealed that under speeded presentation, the darkening manipulation of this study caused a performance reduction on all three tasks relative to the original images of Study 1: Role ($t(40.92) = 2.24, p_{unc} = .030, p_{cor} = .060, d = 0.68, BF_{01} = 0.48$), Harm ($t(32.97) = 3.94, p_{unc} < .001, p_{cor} < .001, d = 1.20, BF_{01}$...
= 0.008), and Moral Wrongness ($t(78.72) = 4.61, p_{unc} < .001, p_{cor} < .001, d = 1.00, BF_{01} = 0.001$).

Crucially, and in line with our predictions, although the darkened images of this study reduced Role performance to similar levels as Study 3 (as there was no difference in Role performance between these two studies, $t(45.65) = 0.44, p_{unc} = .663, p_{cor} = .663, d = 0.13, BF_{01} = 3.26$), the manipulation had a stronger effect on both Harm and Moral Wrongness judgments relative to Study 3 (Harm: $t(49.18) = 3.62, p_{unc} < .001, p_{cor} < .001, d = 0.99, BF_{01} = 0.02$; Moral Wrongness: $t(75.65) = 2.58, p_{unc} = .012, p_{cor} = .012, d = 0.57, BF_{01} = 0.34$). In other words, the darkening manipulation of this study impaired harm and moral wrongness performance more so than the patient role manipulation of Study 3. Thus, Harm was affected somewhat independently of Role extraction, likely leading to the corresponding deficits in accurately assigning Moral Wrongness to the agents in the interaction.

**FIGURE 7**

RESULTS FOR STUDY 4
NOTE.— (a) In Study 4, the luminance level of the original Study 1 images was decreased to a level of 20%. This manipulation was expected to disrupt Harm extraction, and lead to concomitant disruptions to Moral Wrongness extraction. (b) Results from Study 4 are plotted alongside data from Studies 1 and 3, with the same social interactions due to unspeeded norming excluded across all plotted data (see Study 4 methods for details). Points are mean d-primes across participants for each judgment condition and error bars are 95% confidence intervals. Horizontal lines above conditions reflect pairwise significance tests comparing data from Study 4 to Study 1 and Study 3 (corrected for two planned between-study comparisons within each task, using the Bonferroni-Holm method). Relative to Study 1, Study 4 showed reductions in all judgment conditions. Crucially, although the darkening manipulation of Study 4 reduced role extraction to the levels of Study 3, it had a stronger effect on both Harm extraction and Moral Wrongness assignment than did Study 3; thus, Harm was affected somewhat independently of Role extraction, and led to corresponding decrements in the accuracy of moral judgments. *** $p < .001$; * $p < .05$; † $p = .060$; n.s. not significant, $p = .663$.

Discussion

Darkening the original images from Study 1 impaired performance on all tasks. Relative to Study 3, however, the impairment of role detection was comparable even as the impairments of harm and moral wrongness ascriptions were greater, in line with our hypotheses. Thus, making moral judgments based on thin slices may rely on extracting harm and role features that
are independent, or at least partially so.

**GENERAL DISCUSSION**

We find that participants can make accurate thin-slice moral judgments of social interactions occurring between people. Supporting our hypotheses, Study 1 used a controlled image set and design to show that people are indeed capable of moral thin-slicing. Study 2 presented these same images at various durations and found that people can only make accurate moral judgments once they have accurately extracted role and harm information, which they do after different durations of viewing time. Studies 3-4 made role and harm features harder to detect and found concomitant detriments to the accuracy of moral judgments, suggesting that harm and role information affect moral judgments in at least a partially, if not completely, independent manner.

Somewhat unexpectedly, we also found that participants tend to blame the patient (or victim) of a harmful interaction regardless of whether they image is shown briefly. To our knowledge, this is the first demonstration of victim blaming for speeded presentations of generic moral interactions, suggesting that victim blaming may the default way in which people categorize harmful social interactions.

How Is Moral Thin-Slicing So Fast?

Given that people are more accurate at snap moral judgments than one would think, how is this possible? The current work adds to the literature on how the mind computes moral
judgments, by suggesting that such judgments do not have to be slow and effortful; rather, the human visual system in some cases rapidly extracts the high-level information on which moral judgment depends, such as role and harm. Furthermore, the visual system not only extracts such information, which previous literature in some cases has provided evidence for (Hafri et al., 2013; 2018; De Freitas & Alvarez, 2018), but it integrates these moral ‘atoms’ such that they inform moral judgments about events viewed at a brief glance. Notably, this integration was not a given, as there are many cases in other areas of psychology where disparate sources of visual or spatial information fail to be integrated towards a common behavioral goal (e.g., for grasping an object, or reorienting in an unfamiliar environment; Rossetti, 1998; Hermer-Vasquez et al., 1999).

Of course, despite the ability to make moral judgments quickly from a brief glance, this does not mean that people do not sometimes slowly deliberate over whether an event was causal, harmful, and so forth, which thought experiments like the trolley problem clearly illustrate (although such scenarios are overly contrived, and deliberately designed to stump readers; De Freitas, Anthony, Censi, & Alvarez, 2020; De Freitas et al., 2021). Yet the current results suggest that the visual system helps produce a rapid moral judgment when confronted with a range of typical social interactions, circumventing the need to deliberatively mull over this information.

As such, these findings stand in contrast to the characterization of moral judgment as reliant on purely rational inferences about inputs such as causation, harm, etc. without substantive contribution from sensory processing (Martinez & Jaeger, 2016; Olson, McFerran, Morales, & Dahl, 2016; Xie, Yu, Zhou, Sedikides, & Vohs, 2014). These characterizations suggest that visual processing is involved in moral judgment only in a rudimentary sense, e.g., to recognize colors or objects, and their spatial locations within an image.
Our results are also related to, yet distinct, from the social intuitionist model of moral judgment (Haidt, 2001). At a broad level, our results are in line with the idea of that model that moral judgments can be fast, because they often rely on fast mental processes. Yet whereas the social intuitionist model finds that moral judgments are fast once a basic understanding of the scene has already been constructed (e.g., one already knows the event roles and degree of harm), the current work finds that moral judgments are also fast because the mind is fast at understanding the scene in the first place; at least in the case of visual scenes, visual processes rapidly extract the high-level ‘atoms’ of moral judgment, and the mind then rapidly integrates them to produce a moral judgment.

Of course, this is not to say that moral judgments are not also influenced by information that goes well beyond sensory input, such as contextual information about the social interaction, levels of arousal (J. D. Greene, Sommerville, Nystrom, Darley, & Cohen, 2001), subjective values (Newman, De Freitas, & Knobe, 2015), or various heuristics and biases (De Freitas & Johnson, 2018; Haidt et al., 1993). Moreover, our results do not explain why people care about making moral judgment in the first place, nor how they know the moral rules for how to combine pieces of information in order to make accurate moral judgments (see Curry, Mullins, & Whitehouse, 2019; De Freitas, Thomas, DeScioli, & Pinker, 2019).

Is Moral Thin-Slicing Accurate?

In one sense, we can say that moral thin-slicing is accurate in that people’s speeded judgments match their unspeeded judgments for some properties (like event roles) and are a close match for harm and moral judgment. Yet we can still ask whether they are accurate relative
to more objective measures. After all, just because people show high speeded-unspeeded agreement does not mean that these are good evaluations of the true (but hidden) moral qualities of the individuals depicted in an image. As an example, people also agree on the trustworthiness of faces (Todorov, Pakrashi, & Oosterhof, 2009), but this does not necessarily mean they are correct. Does moral thin-slicing invite similar concerns?

We believe the answer is mixed. On the one hand, features indicative of role and harm are likely to be very reliable reflections of the interaction occurring in an image. At an immediate level, then, if it looks like one agent is harming another, it is likely that this is indeed what is happening. On the other hand, whether someone is truly morally wrong depends on a host of factors other than the social interaction that is immediately occurring. For instance, the agent in the immediate interaction may have justifiable reasons for harming the patient, such as self-defense or because they fear that the patient might harm another person, thereby making them less morally wrong than the immediate interaction suggests. Without further contextual information, it is possible that judging an actor’s moral wrongness based on an immediate interaction will be globally inaccurate, given all considerations.

Relatedly, in the current studies we found that, even under speeded presentation, people judged that patients of harmful interactions were more morally wrong than patients of unharmful interactions: a form of ‘victim blaming’. This response pattern might be useful in aggregate. After all, it may be generally true that people do not harm others for no reason at all, causally implicating the victim to some extent. At the same time, these inferences can be inaccurate or even pernicious in individual instances. After all, an agent may also harm for no reason, and their reasons for harm may provide insufficient justification for the harms they cause.

Importantly, our results suggest that victim blaming may be more challenging to
eradicate than has typically been suggested by studies that focus on the role of individual or demographic drivers of victim blaming. Instead of targeting such differences (e.g., people with certain political beliefs), we may need to target the widely shared cognitive factors that drive the bias. By the same token, if victim blaming is a more universal categorization bias, then victims may need to be more wary of sharing news of their victimization online, as in posts related to causes such as Black Lives Matter, #MeToo and #TimesUp (Mendes, Ringrose, & Keller, 2018; Merkin, 2018). Even if social media users share such posts with seemingly like-minded users, they may be met with a mixture of support and attack. Future work should find ways to (i) reduce incidences of inaccurate victim blaming, so that (ii) victims find help for and raise awareness of transgressions, and (iii) transgressors, whether companies or other people are held accountable.

Limitations

While we took steps to make our stimuli as controlled as possible, real-world online moral transgressions also involve pairings between images and text (e.g., social media posts), distractions (e.g., music playing in the background), varying contexts (e.g., news sources with varying credibility), and posts whose influence has as much to do with the content itself as with how the content is edited to selectively include and exclude information. Future work should explore creative ways to test moral thin-slicing in the field.

Additionally, while we took various steps to present images for precise, rapid millisecond durations, we did not visually mask images after presentation. This was intentional, since no single mask could have equally masked relevant features across all four of our tasks, which
differed in various ways (e.g., a simple color feature vs. a configural role feature). For instance, if we had created scrambled stimuli as a mask, then there would be no way of ensuring that the exact same scrambled stimuli were equally effective at masking the different types of visual features in the stimuli, e.g., perhaps the colors in the scrambled image would more effectively mask colors in the images than the body parts in the scrambled image would mask the role information. Thus, any single mask could have artificially introduced differences in the difficulty of the different tasks. A drawback of this methodological choice is that, although visual processing likely played a crucial role in enabling speedy moral judgments, we cannot say that our rapid presentations isolated feedforward visual processing without also being affected by top-down processing typical of recurrent, attentional, and cognitive processes (Chikkerur, Serre, Tan, & Poggio, 2010; Coltheart, 1980; P. M. Milner, 1974).

Conclusion

Moral transgressions are perhaps more accessible than ever, and people view such content at a rapid rate online. Existing work suggests that people make fast moral impressions once they already know what has transpired, e.g., who did what to whom, and whether there was harm (Haidt, 2001). Here, we find that people are also fast at extracting the atoms of moral judgment from a visual scene in the first place and integrating them to decide who is morally wrong, doing so for scenes presented within the blink of an eye (100 ms). In short, people’s moral judgments can keep pace with the breakneck speeds at which they view moral transgressions online; indeed, their online viewing speeds may reflect this very ability.
Author Note

All data and analysis code have been made publicly available on Github and can be accessed at [persistent URL or DOI].

REFERENCES


