A Computational Theory of Dignity

Kartik Chandra[1], **Joshua B. Tenenbaum**[2] **& Rebecca Saxe**[2]
[1]Computer Science and Artificial Intelligence Laboratory, MIT
[2]Department of Brain and Cognitive Sciences, MIT

Abstract

People seem to hold a variety of conflicting intuitions about the concept of dignity. Here, we seek to reverse-engineer the computational basis of these intuitions. We propose a Bayesian model that casts intuitions about dignity as computations about what agents' actions imply about their own and others' social rank. We show through two behavioral experiments that our model captures people's intuitions well, reconciling seemingly conflicting intuitions with a common computational framework.
Keywords: theory of mind; computational models

Introduction

'What do you think dignity's all about?'

The directness of the inquiry did, I admit, take me rather by surprise. 'It's rather a hard thing to explain in a few words, sir,' I said. 'But I suspect it comes down to not removing one's clothing in public.' —Ishiguro, *The Remains of the Day*

Our intuitions about dignity permeate nearly every facet of our social lives: how we stand, eat, and dress, how we treat professors, presidents, and prisoners, and how we conduct graduations, weddings, and funerals. Doctors appeal to dignity when treating patients (Beyleveld & Brownsword, 2001), lawyers when defending clients (Luban, 2005), and leaders when asserting human rights (McCrudden, 2008). Claims about the inviolability of human dignity are invoked in the preambles to the constitutions of Germany and India, in landmark legal decisions like *Miranda v. Arizona*, and in key resolutions of international law, such as the U. N. Universal Declaration of Human Rights. In this and many other ways, dignity appears to be a fundamental social concept.

But what exactly *is* dignity? Philosophers have long struggled to make sense of the many seemingly-irreconcilable intuitions that people have about dignity. For example, consider the following intuitions about the origins, properties, and implications of dignity:

1. Rank: High social status confers dignity on office-bearers, imposing standards of proper treatment from lower-status individuals (Waldron, 2012).
2. Equality: Dignity is a matter of equal treatment. The United Nations' Universal Declaration of Human Rights is founded on the "recognition of the inherent dignity and of the equal and inalienable rights of all members of the human family."
3. Autonomy: Dignity derives from the human capacity and desire for agency (Kant, 1785; Macklin, 2003; Pico della Mirandola, 1496; Pinker, 2008). A prisoner in shackles, or an elderly person unable to use the bathroom unassisted, suffers an indignity by being deprived of their autonomy.
4. Universality: *All* humans have dignity, regardless of agency. For example, the dignity of babies, people with severe cognitive disabilities, and even of dead bodies is invoked in debates on abortion, medical ethics, and museum displays of human remains (Killmister, 2017; Scarre, 2003; Watson, 2016).
5. Labor: Receiving fair exchange and recognition for one's labor is a matter of dignity (Engelmann & Tomasello, 2019; King Jr, 2012).
6. Comportment: Some ways of holding, moving and clothing one's body constitute dignified behavior (Martin & Martin, 2010; McClelland, 2011; Waldron, 2017). Upright posture and restrained bodily actions, like "holding one's head high," are signatures of dignity. Some natural actions of the body, like defecating or sweating, are a challenge to dignity.

It is not clear how to reconcile these intuitions. Equality and Rank are in tension, Universality contradicts Autonomy, and Labor seems conceptually unrelated to Comportment.

To illustrate the tension in people's intuitive qualitative descriptions of dignity, consider the following data (collected as part of Experiments 1 and 2, described below). We asked $N = 482$ participants to rate the degree to which they agree or disagree with the following statements, which endorsed either Rank-based or Equality-based conceptions of dignity:

R1 Dignified treatment of a person with high social rank (e.g. judges, professors, doctors) shows that you consider their status to be higher than your own.

R2 Dignified treatment of a person means treating them appropriately according to whether their status is higher or lower than your own.

R3 The way to treat a person with dignity varies, depending on that person's rank or social status.

E1 Dignified treatment of a person shows that you consider them as an equal to yourself.

E2 Treating someone with dignity means treating them the same as all other people, regardless of their social rank or status.

E3 The way to treat a person with dignity does not depend at all on that person's rank or social status.

Statements R1/E1, R2/E2, and R3/E3 are pairwise contradictory. Nonetheless, we preregistered the prediction that all six statements are aspects of the concept of dignity, and so many individual participants will agree with both E1 and R1, E2 and R2, and/or E3 and R3. Indeed, 81% of participants agreed

In G. Boleda, I. Dautriche, M. Maldonado, C. Saldana and T. Torrent (Eds.), *Proceedings of the 48th Annual Conference of the Cognitive Science Society*

with E1 and 46% with R1 (37% with both); 81% agreed with E2 and 51% with R2 (39% with both); and 73% agreed with E3 and 30% with R3 (12% with both).

Facing such contradictions, many philosophers have concluded that "dignity" refers to multiple distinct concepts (Nordenfelt, 2004; Rao, 2011; Schroeder, 2008), or that it is an "essentially contested concept" (Rodriguez, 2015) used as a "placeholder" for anyone to "insert their own theory" where convenient (McCrudden, 2008). Others dismiss dignity as a "vacuous" (Bagaric & Allan, 2006), "useless" (Macklin, 2003) or "stupid" (Pinker, 2008) concept.

In this paper, we take a computational approach to reverse-engineer what people are computing when they report that someone is acting with dignity, or treating someone else with dignity. We begin by identifying a key feature of social situations to which people are likely to be highly sensitive: implications about agents' social ranks generated by their actions. We propose that implications about rank constitute one source of intuitions about dignity. We formalize this hypothesis by building a computational model of people's intuitions about dignity, and we test our model with a behavioral experiment, showing that it explains people's intuitions about dignity better than competing models that operationalize alternate theories from the literature. Finally, by closely analyzing the data from this experiment, we identify two factors that were missing from our initial hypothesis. We formalize these refinements in our computational model, and with a second behavioral experiment show that these refinements better explain people's intuitions. Our rank-based theory of dignity best predicts data from even those participants who explicitly self-report that dignity is a matter of equal treatment. (Preregistrations: https://aspredicted.org/s3jy8i.pdf, https://aspredicted.org/2xw96y.pdf.)

Our proposal: implications about social rank

To understand intuitions about dignity, we begin with the assumption that one of the fundamental computational challenges of social cognition is coordinating groups of agents with hidden mental states, large action spaces, and long planning horizons. As Jara-Ettinger and Dunham (2024) argue, one way we make this computation tractable is through the abstraction of *roles:* bundles of rights, responsibilities, and norms assigned to agents in a community. Like positions on a basketball team or lanes on a highway, roles ease the cognitive burden of coordination. But not all roles in an institution are equally desirable. Agents may prefer to occupy particular roles because of the powers and privileges those roles bestow on their bearers. Here we call roles that are equipped with an established social hierarchy, *ranks*.

Ranks are social fictions that are assigned by consensus. Typically, what it means to occupy a high rank (or desirable role) and enjoy its associated powers is for others to collectively agree that you occupy that rank and agree to abide by the associated norms (Searle, 1995, 2010). For example, King Lear was denied the powers of his rank by his daughters, who refused to abide by the norms associated with his royal status.

Importantly, collective agreement about ranks is not a directly observable property of the physical world. Like any mental state, its presence can only be inferred indirectly from people's observable behavior. Hence, people live under constant uncertainty about whether or not they truly occupy the rank they think or hope that they do. In the face of this uncertainty, we should expect individuals (1) to *claim* desired ranks through communicative actions, (2) to value having claims of high rank *affirmed* by others (McCall & Simmons, 1966), (3) to find it costly to have claims of high rank *denied* by others (Torres & Bergner, 2010), and (4) to be averse to signals that others perceive them in an undesirable or low rank. We propose that this sensitivity to rank-related implications is one source of our intuitions about dignity. In particular, *acting with dignity includes signaling your own high rank (or obscuring your low rank), and treating someone with dignity includes signaling their higher rank (or obscuring their lower rank).*

Consider for example the case of the Great French Mustache Strike of 1907. When the first modern restaurants were established in Paris, their function was in part to offer diners a taste of the noble elite lifestyle, including being waited on by domestic servants. Domestic servants at the time were forbidden from wearing mustaches, as a sign of their lower rank; hence, Parisian waiters were required to shave their mustaches. The waiters however found this requirement humiliating and an affront to their dignity, ultimately taking to the streets in protest and shutting down the French dining scene until they won back the right to wear mustaches. Consider a Parisian restaurant owner who wore luxuriant mustaches while insisting that his male employees be clean shaven. Our hypothesis predicts that while his own action and comportment may be dignified, he is not treating his waiters with dignity. On the other hand, an owner who shaved his mustache to emphasize equality with the waiters would be treating them with dignity.

Empirical approach To test this theory, we conducted two behavioral experiments. In each experiment, participants watched short videos depicting simple social interactions in a hypothetical society with a known set of ranks and rank-associated norms. (We used hypothetical societies, with unfamiliar though plausible norms, so that we could control people's priors and rely on computation rather than memory.) The videos vary what the interactions revealed about the agents' ranks. Participants reported their intuitions about dignity in these scenarios, which we compared to the predictions of a computational model instantiating our theory, as well as alternate models that instantiate pure-equality and pure-autonomy accounts of dignity.

Experiment 1

Procedure We recruited 120 participants on Prolific. Participants were paid $11.25 for an estimated 45 minute study (median time was ultimately 43 minutes). Participants completed quizzes and attention checks to ensure that they under-

stood the paradigm, and the details of each scenario. Finally, participants completed a post-survey asking them to report the degree to which they agreed with statements E1-3/R-3 (data reported in the introduction, aggregated with participants from Experiment 2). Following our preregistered exclusion criteria, we retained data from $N = 82$ participants.

Stimuli Our experimental paradigm probed intuitions about dignity in simple social interactions, independently manipulating societal norms, relative ranks of agents, and the actions those agents take for themselves and toward others. Participants were told about a faraway village with two social ranks: *peasant* (the low rank) and *ambassador* (the high rank). Most villagers are peasants, but through exceptional community service some can be promoted to the desirable and powerful rank of ambassador. Participants were told that all villagers dress the same regardless of rank, and that their ranks cannot be determined on the basis of appearance.

Next, each participant was familiarized with one of two possible societal norms dictating the type of dish from which villagers are permitted to drink. Half of the participants learned the *privilege* norm: only ambassadors may use glass mugs, while both peasants and ambassadors may use metal mugs. The other half learned the *relegation* norm: ambassadors only drink from glass mugs, while peasants may drink from both glass and metal mugs. All participants were told that the norm is strictly observed by all villagers, with no exceptions.

Finally, participants were shown a series of 10-second videos. Each video depicted two agents of different ranks seated in a tavern: an ambassador and a peasant. The agent seated closer to the bar (who could be either the ambassador or the peasant) gets up to bring drinks for the table. They pick two mugs from the bar and place one in front of themselves and one in front of the other agent. After each video finished, participants reported whether they thought the actor "treated [the target] with dignity" and whether they thought the actor "acted with dignity themselves," both on 5-point Likert scales ranging from "strongly disagree" to "strongly agree."

We exhaustively manipulated the following variables across the videos: (A) whether participants knew the ranks of the actor and target, or whether the agents' ranks were unknown, (B) which type of mug the actor chooses for themself, and which type they choose for the target (subject to the constraints of the norm in effect), and (C) the relative costs of acquiring each type of mug, by varying whether the two mugs were at equal distance, or one was near and the other far along the bar, requiring the actor to go out of their way to fetch the far mug. Each participant saw all 21 possible scenarios for their assigned norm condition (Fig 1), in randomly shuffled order. Aggregating across both norms, we collected data on a total of $21 \times 2 = 42$ scenarios. We collected two dependent variables per scenario (judgments of acting with dignity and treating with dignity), for a total of $42 \times 2 = 84$ dependent variables.

For example, imagine a scenario in the privilege norm condition (Fig 1). The actor is known to be the ambassador. The glass mugs are near the actor, and the metal mugs are far away. The actor picks a glass mug for themselves, but then goes out of their way to pick a metal mug for the target (the peasant). We predict that people will say that the actor did *not* treat the target with dignity, because the actor went out of their way to emphasize the target's lower rank with a metal mug. However, the actor acted with dignity, because their own glass mug signals their high rank.

Main model Consider a society with a set R of possible roles. Agents in this society can take actions from the set A of possible actions, but allowable actions are constrained by a role-dependent norm n, modeled as a binary function. In particular, an actor with role $\rho_{\text{actor}} \in R$ can only take action $a \in A$ towards a target agent with role $\rho_{\text{target}} \in R$ if $n(a, \rho_{\text{actor}}, \rho_{\text{target}}) = 1$. (We use $\vec{\rho}$ as shorthand for the pair $\rho_{\text{actor}}, \rho_{\text{target}}$ below.) We model an observer's intuitions about whether an actor acted with dignity, and whether they treated a target with dignity, by constructing a Bayesian inverse-planning model in the spirit of Rational Speech Acts (Frank & Goodman, 2012) and Rational Communicative Social Actions (Radkani et al., 2025). Our model consists of a hierarchy of actors and observers who reason recursively about each other.

Let us say that a dignity-naïve actor (at "level 0") takes actions by considering only the action cost $c(a)$ and permissibility under norm n. We will model such an actor as softmax-rational (Luce et al., 1959): $p_0(a \mid \vec{\rho}) \propto \exp(\beta \cdot c(a)) \cdot n(a, \vec{\rho})$, where the parameter β modulates the noisiness of the actor's choice. A dignity-naïve observer (at "level 1") observes an actor's action a and seeks to infer the roles of the agents from this action. We can model this observer's inference via Bayes' rule, $p_1(\vec{\rho} \mid a) \propto p_0(a \mid \vec{\rho}) p_{\text{prior}}(\vec{\rho})$, where p_{prior} is a prior over roles of unknown agents (which we set to uniform).

Now, consider a dignity-sensitive actor (at "level 2") who is concerned not only with taking a low-cost norm-constrained action, but also about the role-related implications of that action. Recall that on our account, acting with dignity means signaling your own high rank (or obscuring your low rank), and treating someone with dignity means signaling that their rank is higher (or obscuring that their rank is lower). To formalize these glosses, suppose that each role has some value $V(\rho)$ associated with it, such that $V(\rho)$ is higher for more desirable roles, effectively creating a ranking of roles. The dignity-sensitive actor places weight α_{actor} on the expected utility of their role computed by a dignity-naïve observer, and places weight α_{target} on the delta between the expected utility of the target's role and the actor's role, again as computed by a dignity-naïve observer. This can be modeled as $p_2(a \mid \vec{\rho}, \vec{\alpha}) \propto \exp(\beta \cdot U) \cdot n(a, \vec{\rho})$, where

$$U = c(a) + \alpha_{\text{actor}} \cdot E_{\vec{\rho} \sim p_1(\vec{\rho} \mid a)}[V(\rho_{\text{actor}})] + \alpha_{\text{target}} \cdot E_{\vec{\rho} \sim p_1(\vec{\rho} \mid a)}[V(\rho_{\text{target}}) - V(\rho_{\text{actor}})].$$

Finally, the dignity-sensitive observer (at "level 3") infers the actor's weights $\alpha_{\text{actor, target}}$ in order to report judgements about dignity. In particular, the dignity-sensitive observer infers $p_3(\vec{\alpha}, \vec{\rho} \mid a) \propto p_2(a \mid \vec{\rho}, \vec{\alpha}) p_{\text{prior}}(\vec{\rho}, \vec{\alpha})$, where p_{prior} is now expanded to include a prior over $\vec{\alpha}$ as well (we set the prior on $\vec{\alpha}$ to be unit Gaussian). Then, the dignity-sensitive

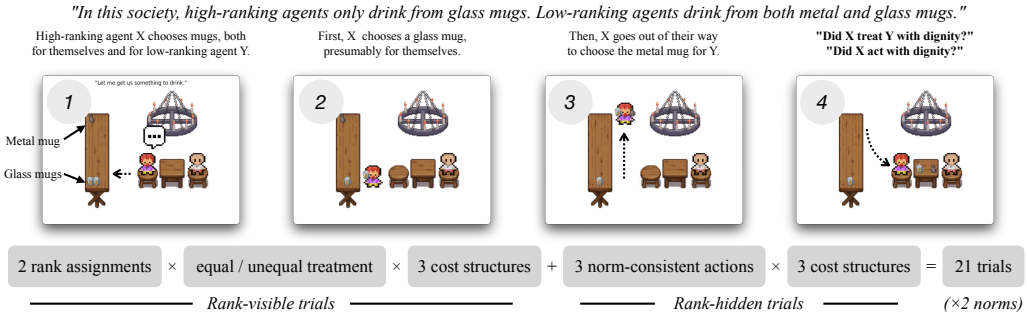

Figure 1: Example stimulus from our experimental paradigm, and the full stimulus set. See text for description.

observer computes the degree to which the actor acted with dignity as $E_{\vec{\alpha},\vec{\rho}\sim p_3(\vec{\alpha},\vec{\rho}|a)}[\alpha_{\text{actor}}]$, and the degree to which the actor treated the target with dignity as $E_{\vec{\alpha},\vec{\rho}\sim p_3(\vec{\alpha},\vec{\rho}|a)}[\alpha_{\text{target}}]$.

This model has two free parameters: c and V, parametrized as the marginal cost of acquiring the "far" mug and the marginal value of the high rank over the low rank. (Without loss of generality, we set $\beta = 1$.)

Alternate models We considered two competing models, M2 and M3. M2 instantiated the hypothesis that dignity is a matter of equal treatment: to treat someone with dignity is to create an equal outcome for yourself and them, and to act with dignity means to show that you value equal treatment. Under this model, the dignity-sensitive actor considers the cost of the action $c(a)$, and also whether the action creates equal outcomes for the actor and the target. This term is given by $\alpha_{\text{equality}}\mathbf{1}_{o_{\text{actor}}=o_{\text{target}}}$ where $\vec{o}$ are the outcomes that a creates for the target and actor. The dignity-sensitive observer infers the actor's α_{equality}. This observer reports that the degree to which the actor treated the target with dignity is proportional to $\mathbf{1}_{o_{\text{actor}}=o_{\text{target}}}$, and the degree to which the actor acted with dignity is proportional to the inferred α_{equality}.

M3 sought to instantiate the hypothesis that dignity is a matter of autonomy. One way to operationalize autonomy in our paradigm is to say that choosing the mugs always carries more autonomy (and thus dignity) than passively receiving a mug. This account predicts that actors have more dignity than recipients, with no additional variance across our stimuli. For a non-trivial alternative model, we operationalized autonomy as preference satisfaction because in these simple interactions, autonomy would typically enable preference satisfaction. In M3, to treat someone with dignity is to give them what they want, and to act with dignity is to get what you want. Under this model, the dignity-sensitive actor considers the cost of the action $c(a)$, and also whether the actor and the target get what they want. These are given by $\alpha_{\text{actor}}\mathbf{1}_{o_{\text{actor}}=g_{\text{actor}}}$ and $\alpha_{\text{target}}\mathbf{1}_{o_{\text{target}}=g_{\text{target}}}$, where $g_{\text{actor, target}}$ are the actor and target's goals, respectively (assumed to be known to the actor). The dignity-sensitive observer infers agents' goals $\vec{g}$ by Bayesian Inverse Planning and then infers $\vec{\alpha}$ in order to report judgments of acting/treating with dignity as in M1.

We implemented these models using the memo programming language (Chandra et al., 2025).

Results

Overall model fit First, we fit M1 to our 84 data points, minimizing mean squared error between model predictions and average participant responses. M1 predicted $R^2 = 0.62$ of the variance in the aggregate data (Figure 2, left).

Tests of generalization Next, we performed the following tests of generalization:

1. Generalization across stimuli: We randomly partitioned our 84 data points into 7 groups of 12 data points each and performed a 7-fold cross-validation by testing M1 on each of the 7 groups by fitting its parameters to the remaining 72 data points. M1 predicted $\mu = 0.57, \sigma = 0.13$ of variance. To establish a baseline, we repeated this analysis with the labels shuffled (randomly permuted), and the same model explained only $\mu = 0.05, \sigma = 0.03$ of the variance.
2. Generalization to unseen actions: For a given norm, there are three possible norm-consistent actions. For each of the three actions, we computed the percent variance explained among data points, fitting M1 to data from the remaining two actions. Across these 6 fits, M1 explained $\mu = 0.45, \sigma = 0.37$ of the variance.
3. Generalization across rank visibility: We fit M1 to trials where the agents' ranks were visible, and predicted data from trials where agents' ranks were hidden ($R^2 = 0.66$). Similarly, we fit M1 to trials where agents' ranks were hidden, and predicted data from trials where agents' ranks were visible ($R^2 = 0.58$).
4. Generalization to new norms: We fit M1 to data from the *privilege* norm, and predicted data from the *relegation* norm ($R^2 = 0.68$), and vice versa ($R^2 = 0.25$).

Model comparison We tested M1 against our alternate models M2 (equality) and M3 (agency). We repeated the 7-fold cross-validation with M2 and M3 and compared the models' R^2 on the held-out test sets. (We compared generalization, rather than overall model fit, to make sure that M1 did not succeed merely due to its higher parameter count leading to overfitting.) M2 explained $\mu = 0.05, \sigma = 0.04$ of the variance, significantly less than M1 ($p < 0.05$ by paired t-test). M3 explained $\mu = 0.29, \sigma = 0.21$ of the variance, also less than M1 ($p < 0.05$). These results support our account of

dignity as including communication about social rank, rather than just equal treatment or agency.

Interim discussion These results suggest that our rank-based theory of dignity predicts people's intuitions well. M1 captures a large amount of variance in people's judgments, generalizes robustly, and performs better than alternate models. But there is still significant variance left to be explained. Inspecting the deviations between M1 and our data led us to form two new hypotheses:

A When treating others with dignity, people's utility depends asymmetrically on the implied rank difference. Avoiding making the target appear *lesser* than the actor is more important than making the target appear *greater* than the actor.

B One component of acting with dignity is appearing to treat others with dignity.

To test these hypotheses, we ran a second experiment.

Experiment 2

Procedure Experiment 2 followed the same paradigm and procedure as Experiment 1, except that to facilitate within-subjects analyses, each participant saw *both* norms. Trials were grouped by norm, and the two norms were presented in randomized order. To minimize confusion between the norms, we duplicated all stimuli to replace mugs with bowls. Participants saw mug stimuli for one norm, and bowl stimuli for the other norm. We collected data from $N = 400$ post-exclusion participants on Prolific. Participants were paid $15 for an estimated 1 hour study (median time 1:03).

Models We formalized hypotheses A and B as extensions of model M1. To formalize A, we revised the level-2 actor's utility to include a new asymmetric term, $a \cdot E[\min(0, V(\rho_{\text{target}}) - V(\rho_{\text{actor}}))]$. The parameter a modulates the degree to which the actor weighs this term. To formalize B, we added a new layer of recursion: we considered a level-4 actor whose utility includes an additional term for convincing the level-3 observer that he is treating the target with dignity. The level-5 observer infers this actor's weight on this term, and the new parameter b modulates the degree to which the observer weighs this inferred weight when judging if the actor acted with dignity. The new M4 includes both updates. M4a includes only A and M4b includes only B.

Results

Overall model fit Across the 2 norms, 21 vignettes per norm, and 2 dependent variables per vignette, we had a total of 84 data points from each participant. We first fit M4 to the 84 data points, averaging responses across subjects, and minimizing mean squared error between model predictions and average participant responses. M4 predicted $R^2 = 0.85$ of the variance in the aggregate data (Figure 2, right).

M4 showed high generalization on the same measures as in Experiment 1 (generalization across stimuli, to unseen actions, across rank visibility, and to new norms). Additionally, we tested generalization to new participants. When fit to Exp. 1 data, M4 predicted $R^2 = 0.69$ of the variance in Exp. 2 data; when fit to Exp. 2 data, M4 predicted $R^2 = 0.65$ of the variance in Exp. 1 data. For comparison, in a direct regression Exp. 2 data explained $R^2 = 0.76$ of variance in Exp. 1 data, and vice versa. This noise ceiling gives an upper bound on the generalization we should expect across these experiments.

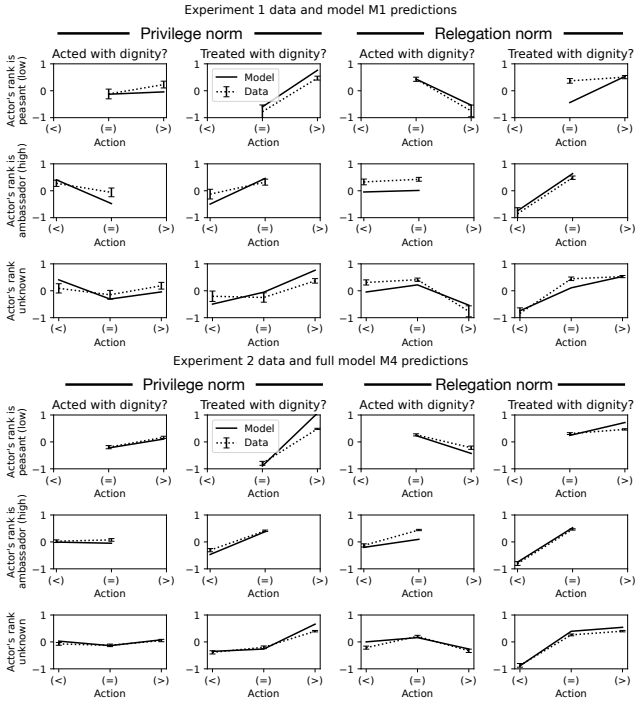

Figure 2: Data from Experiments 1 and 2, aggregating over the manipulation of action cost. Each plot varies whether the actor treated the target equal-to (=), less-than (<), or greater-than (>) themselves (subject to norm permissibility).

Model comparisons We compared M4 to competing models M2 (equality) and M3 (agency). In a 7-fold cross-validation, M2 predicted $\mu = 0.05, \sigma = 0.06$ of the variance, significantly less than M4 did ($p \ll 0.01$). M3 predicted $\mu = 0.14, \sigma = 0.12$ of the variance, also significantly less than M4 did ($p \ll 0.01$). We conclude that M4 best explains our data: people's intuitions about dignity are driven by considerations of rank, rather than purely considerations of equality or agency (Fig 3).

Model ablations We compared M4 to our original model M1. On Experiment 2 data M1 predicted $\mu = 0.55, \sigma = 0.14$ of the variance, significantly less than M4 did ($p \ll 0.01$). Adding only asymmetry (M4a) improved the model fit over M1, though not as much as the full model M4 did ($\mu = 0.67, \sigma = 0.08$). Adding only recursion (M4b) also improved the fit over M1, explaining $\mu = 0.83, \sigma = 0.06$ of the variance, as much as the full model M4.

Individual differences Finally, we tested whether participants' differences in responses to this task were aligned with differences in their self-reported conceptions of dignity. In the post-survey (questions E1-3 and R1-3, as discussed in the introduction), some participants more strongly endorsed an equality-based conception of dignity, while others more

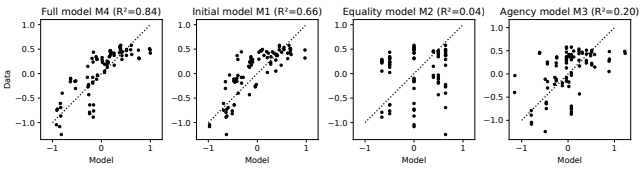

Figure 3: Exp. 2 model comparisons. Full model M4 predicts people's intuitions about dignity better than initial model M1, and alternate models M2 (equality) and M3 (agency).

strongly endorsed a rank-based conception. To capture these differences in self-reported concepts of dignity, we projected each participant's survey responses onto the first principal component of all responses, to get an "E-R score". We then performed two analyses.

First, recall that free parameter a reflects the degree to which the actor weighs signaling that the target is "no less than" themselves, versus signaling the target's true rank. If a participant considers dignity to be a matter of equal treatment, then we should expect a higher fitted value of this parameter a. We fit M4 to each individual participant, and regressed the fitted parameter a against E-R scores. We found no correlation ($r = -0.017, p = 0.73$). Second, we divided participants into quartiles by E-R score. We found that M4 predicts data significantly better than M2 (equality) for participants in each quartile ($p \ll 0.01$ in each quartile). Notably, this was true even in the top quartile, among participants who were most likely to endorse E1-3 over R1-3. Among these participants, M4 explained $\mu = 0.83, \sigma = 0.02$ of variance in responses, while M2 explained only $\mu = 0.009, \sigma = 0.008$ of variance.

We conclude that rather than having two qualitatively different definitions of dignity, people who explicitly report that dignity is a matter of equality and people who explicitly report that dignity is a matter of rank share a single underlying concept of dignity, best captured by M4.

Acting with vs. treating with dignity In a post-hoc analysis, we found that M4 predicted intuitions about *treating* with dignity ($R^2 = 0.92$) better than intuitions about *acting* with dignity ($R^2 = 0.77$). This suggests that there is still more to explain in people's intuitions about acting with dignity.

Discussion

In this paper, we offered the first computational model that quantitatively predicts people's intuitions about dignity. We proposed that these seemingly-conflicting intuitions derive systematically from implications about social rank generated by norm-constrained actions. We formalized our proposal in a computational model and showed through two behavioral experiments that our model qualitatively and quantitatively predicts people's intuitions across dozens of scenarios. This model explains intuitions better than alternative hypotheses that dignity is purely a matter of equality or agency.

In our experiments, we used minimalist stimuli in a hypothetical society, with explicit stipulations about ranks and norms, in order to precisely control people's priors in the simplest possible setting. To generalize our model to naturalistic real-world scenarios, we would have to encode the complex structures of ranks and norms in real societies. This information could be curated by hand, inferred from data, or marshaled on-demand from a large associative store of general world knowledge (Wong et al., 2025). A "stimulus-computable" model of human intuitions about dignity would provide a much stronger test of the generality of our theory. It would also allow us to dispense with explicit stipulations about rank. Our stimuli made rank-related information highly salient, which may have influenced participants to draw on a rank-based conception of dignity. A stimulus-computable model would allow us to use stimuli where rank-related information is implicit, giving a stronger test of our hypothesis.

Our experiments demonstrate how our theory reconciles the seeming conflict between people's intuitions about RANK and EQUALITY, an aspect of dignity related to Goffman (1959)'s notion of "face." However, this approach of formalizing theories in computational models and testing them with behavioral experiments could be applied to test whether other dignity-related intuitions can also be reconciled in terms of implications about rank. For example, we could test the hypothesis that intuitions about fair exchange for LABOR being dignified treatment arise from the implication that the client's relationship with the laborer is cooperative (negotiated among equals) rather than exploitative (dictated by superior). Other intuitions might require additional assumptions to reconcile. For example, intuitions about UNIVERSALITY and COMPORTMENT may depend on a special rank, "human," that ranks above other animals (Pico della Mirandola, 1496; Waldron, 2012), and whose norms apply to all humans equally. Our approach could be used to test the hypothesis that intuitions about universality (e.g. dignified treatment of human remains) and comportment (e.g. "holding one's head high" being a way of acting with dignity) derive from implications about the rank of human.

People's strong desire to be treated with dignity—their willingness to sacrifice material comfort or even endure physical torture over intangible matters of dignity—can at times seem disproportionate and irrational. This seeming irrationality can make it easy to dismiss dignity-related concerns as sentimental, indulgent, or prideful (as Falstaff quips in *Henry IV Part 1*, "Who hath honor? He that died on Wednesday!"). On the contrary, our work suggests that these desires *do* have a rational basis—indeed, that they are a natural consequence of rational social cognition. Roles are resource-rational abstractions that ease the computational challenge of coordinating agents; the desire to occupy and be recognized occupying powerful, high-ranking roles is rational for individuals; and standard rational planning and inverse-planning models capture agents' rank-related decisions and inferences. A rational account of dignity may begin to help us explain and appreciate what motivates people act in the name of dignity, to better understand whether and how to weigh considerations of dignity in high-stakes legal, political, and medical contexts, and to engineer machines that treat people with dignity.

Acknowledgments

We thank the reviewers for their thoughtful feedback on this paper. This work was supported by the MIT Siegel Family Quest for Intelligence, Schmidt Sciences, the Simons Foundation, and the Hertz Foundation.

References

Bagaric, M., & Allan, J. (2006). The vacuous concept of dignity. *Journal of Human Rights*, *5*(2), 257–270.

Beyleveld, D., & Brownsword, R. (2001). *Human dignity in bioethics and biolaw.* Oxford University Press.

Chandra, K., Chen, T., Tenenbaum, J. B., & Ragan-Kelley, J. (2025). A domain-specific probabilistic programming language for reasoning about reasoning (or: a memo on memo). *Proceedings of the ACM on Programming Languages*, *9*(OOPSLA2), 784–814.

Engelmann, J. M., & Tomasello, M. (2019). Children's sense of fairness as equal respect. *Trends in Cognitive Sciences*, *23*(6), 454–463.

Frank, M. C., & Goodman, N. D. (2012). Predicting pragmatic reasoning in language games. *Science*, *336*(6084), 998–998.

Goffman, E. (1959). *The presentation of self in everyday life*. Doubleday.

Jara-Ettinger, J., & Dunham, Y. (2024). The institutional stance. *Under review*.

Kant, I. (1785). *Groundwork for the metaphysics of morals*.

Killmister, S. (2017). Dignity: Personal, social, human. *Philosophical Studies*, *174*, 2063–2082.

King Jr, M. L. (2012). *All labor has dignity* (Vol. 5). Beacon Press.

Luban, D. (2005). Lawyers as upholders of human dignity (when they aren't busy assaulting it). *U. Ill. L. Rev.*, 815.

Luce, R. D., et al. (1959). *Individual choice behavior* (Vol. 4). Wiley New York.

Macklin, R. (2003). *Dignity is a useless concept* (Vol. 327) (No. 7429). British Medical Journal Publishing Group.

Martin, J., & Martin, J. (2010). *Miss manners' guide to a surprisingly dignified wedding*. W. W. Norton & Company.

McCall, G., & Simmons, J. (1966). *Identities and interactions.* Free Press.

McClelland, R. T. (2011). A naturalistic view of human dignity. *The Journal of Mind and Behavior*, 5–48.

McCrudden, C. (2008). Human dignity and judicial interpretation of human rights. *european Journal of international Law*, *19*(4), 655–724.

Nordenfelt, L. (2004). The varieties of dignity. *Health care analysis*, *12*, 69–81.

Pico della Mirandola. (1496). Oration on the dignity of man.

Pinker, S. (2008). The stupidity of dignity. *The new republic*, *28*(05.2008), 28–31.

Radkani, S., Tenenbaum, J. B., & Saxe, R. (2025). What people learn from punishment: A cognitive model. *Proceedings of the National Academy of Sciences*, *122*(32), e2500730122.

Rao, N. (2011). Three concepts of dignity in constitutional law. *Notre Dame L. Rev.*, *86*, 183.

Rodriguez, P.-A. (2015). Human dignity as an essentially contested concept. *Cambridge Review of International Affairs*, *28*(4), 743–756.

Scarre, G. (2003). Archaeology and respect for the dead. *Journal of applied philosophy*, *20*(3), 237–249.

Schroeder, D. (2008). Dignity: Two riddles and four concepts. *Cambridge Quarterly of Healthcare Ethics*, *17*(2), 230–238.

Searle, J. (1995). *The construction of social reality*. Simon and Schuster.

Searle, J. (2010). *Making the social world: The structure of human civilization*. Oxford University Press.

Torres, W. J., & Bergner, R. M. (2010). Humiliation: its nature and consequences. *Journal of the American Academy of Psychiatry and the Law Online*, *38*(2), 195–204.

Waldron, J. (2012). *Dignity, rank and rights.* Oxford University Press.

Waldron, J. (2017). The dignity of old age. *NYU School of Law, Public Law Research Paper*(17-41).

Watson, C. (2016). Womb rentals and baby-selling: does surrogacy undermine the human dignity and rights of the surrogate mother and child? *The New Bioethics*, *22*(3), 212–228.

Wong, L., Collins, K. M., Ying, L., Zhang, C. E., Weller, A., Gerstenberg, T., ... others (2025). Modeling open-world cognition as on-demand synthesis of probabilistic models. *arXiv preprint arXiv:2507.12547*.