CH
U

6

Moral Functions

CHAPTER 6

Imagine designing a machine to mimic the moral judgment of humans. In principle, you may want a machine that is better than humans at making moral judgments. But in practice, that goal may be too farfetched. So, instead, you may want to first make a machine that simply mimics the moral judgment of humans.

The goal of this chapter is to explore the very basics of that machine. To achieve that goal, we will use simple statistical tools that prioritize explicability over accuracy. These tools will help us zoom out of individual scenarios by providing descriptions that are less nuanced, but more generalizable. They will also inform us about the impact of different inputs into moral judgments.

Our exploration will build on the idea of a *moral function*: a mathematical object predicting how people will judge the outcomes of a moral scenario based on inputs, such as who is performing the action, or its level of perceived harm. One input that is of particular interest for us is whether the agent performing the action in a scenario is a human or a machine. Throughout the book, we have seen that people judge human and machine actions differently. This is consistent with the social psychology literature telling us that people judge and punish more severely members of out-groups (in our case machines) than members of the in-group (in our case, humans).[1,*] By using moral functions, we can formalize those differences by exploring how they relate to the characteristics of a scenario.

Our approach will rely on many simplifying assumptions,[†] which we introduce in an effort to prioritize clarity. To make that explicit, we will mention the problems caused by these simplifying assumptions when we introduce them.

To begin, we introduce the *moral space* a quantitative representation of moral judgment. This representation, which we use to abstract away from the details of each scenario, is inspired by Jonathan Haidt's moral foundation theory[2] and is based on three factors: harm, intention, and wrongness. While in principle, we could include many inputs, such as whether the dilemma involves an uncertain outcome or represents

[*] This intergroup bias develops as children grow, and as such, it can be detected as soon as six years old (J. J. Jordan, K. McAuliffe, and F. Warneken, "Development of In-group Favoritism in Children's Third-Party Punishment of Selfishness," PNAS 111 (2014): 12710–12715). Moreover, neuroimaging research shows that people have higher sensitivity (i.e., great activity in the left orbitofrontal cortex) to the suffering of in-group members than out-groups when an out-group member performs the harmful action. (P. Molenberghs, J. Gapp, B. Wang, W. R. Louis, and J. Decety, "Increased Moral Sensitivity for Outgroup Perpetrators Harming Ingroup Members," Cerebral Cortex 26 (2016): 225–233). In an experiment in which Swiss army officers played a prisoner's dilemma, researchers found more cooperation among officers from the same platoon and harsher punishments for defectors from different platoons. (L. Goette, D. Huffman, S. Meier, and M. Sutter, "Group Membership, Competition, and Altruistic Versus Antisocial Punishment: Evidence from Randomly Assigned Army Groups," IZA Discussion Paper No. 5189 (2010), https://papers.ssrn.com/abstract=1682710.) When asked to imagine a theft, undergraduate students assigned higher fines to foreign offenders than to relatives or classmates (D. Lieberman and L. Linke, "The Effect of Social Category on Third Party Punishment," Evolutionary Psychiatry (1 April 2007)). Similar patterns have been observed for affiliations with soccer clubs and political parties, (B. Schiller, T. Baumgartner, and D. Knoch, "Intergroup Bias in Third-Party Punishment Stems from Both Ingroup Favoritism and Outgroup Discrimination," Evolution and Human Behavior 35 (2014): 169–175.) and even among tribes in Papua New Guinea. (H. Bernhard, U. Fischbacher, and E. Fehr, "Parochial Altruism in Humans," Nature 442 (2006): 912–915).

[†] Our presumption is that all the statistical estimates presented here can be improved, but more sophisticated estimation techniques may obscure or distract from the key concepts that we want to communicate.

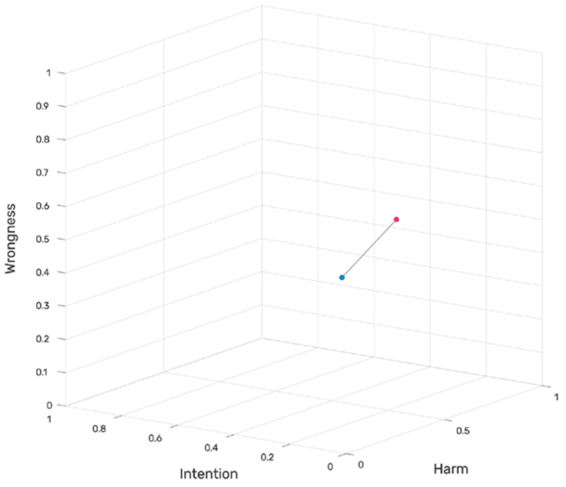
a violation of a moral dimension other than harm, we focus for simplicity only on five variables: the perceived levels of harm, intention, and wrongness of a scenario, and whether the scenario was a treatment or a control (i.e., whether the action was performed by a human or a machine). We then explore how the characteristics of the respondents—the people judging the scenarios—affect moral judgments.

In this representation, each scenario is described by two dots connected by a line. The red dot shows the judgment of the machine action, while the blue dot shows the judgment of same action when conducted by a human. The dots exist in a three-dimensional space defined by wrongness on the vertical axis (the *z*-axis) and harm and intention on the horizontal plane (the *x*- and *y*-axes).

Figure 6.1 illustrates the simplified moral space using average answers for perceived wrongness, harm, and intention. The black line connecting the dots shows that both dots belong to the same scenario. We use a diverging scale for wrongness, meaning that wrongness values range from "Extremely right" (0) to "Extremely wrong" (1), with the neutral value ("Neither wrong nor right") at 0.5. For harm and intention, we use a sequential scale. That is, intention ranges from "Not intentional at all" (0) to

Figure 6.1

Quantitative representation of judgments observed for human and machine actions in a scenario.



"Extremely intentional" (1). Similarly, harm ranges sequentially from "Not harmful at all" (0) to "Extremely harmful" (1).

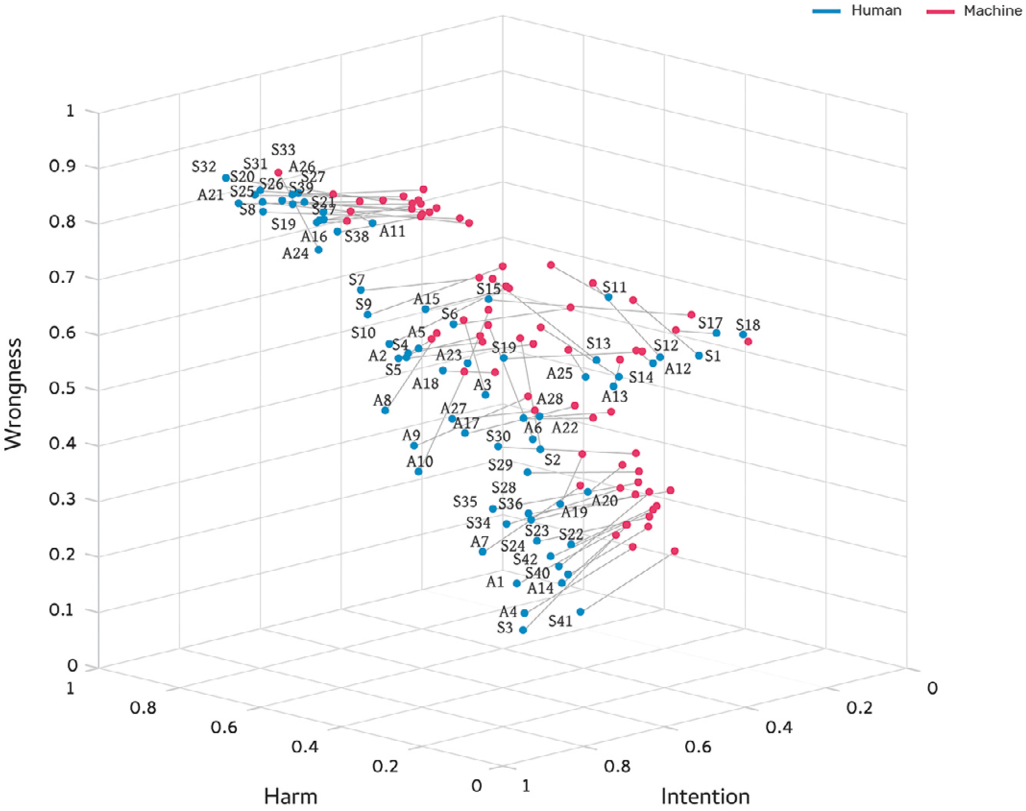
We can use this representation to summarize the patterns found across all the scenarios that included questions on perceived wrongness, harm, and intention. This excludes the privacy and labor displacement scenarios, which did not include these three questions.

Figure 6.2 shows a summary of our experimental results. Note that the moral space is purely descriptive, which allows us to consider wrongness, harm, and intention simultaneously, even though these are all affected by the treatment.

The first finding, which is interesting but slightly obvious, is that moral judgments do not populate the whole space. They fall within a plane that extends from the upper-left corner, with high levels of harm, wrongness, and intention, to the right side of the cube, which shows scenarios with low levels of wrongness and harm. This is because some corners, such as scenarios with no intention or harm, cannot be high in wrongness. Similarly, scenarios high in harm and intention cannot be rated as low in wrongness.

These constraints limit the observation to relatively narrow moral planes. In the next section, we will model these planes mathematically. In this section, we explore the patterns found in this three-dimensional space by looking at the three faces of the cube separately.

Figure 6.3 zooms into the harm-intention plane. Here, we see that machine actions are seen as less intentional than human actions when the level of human intention is relatively high, which is true for most cases in our sample. However, we find six scenarios in which the actions of machines are seen as more intentional than those of people. These six cases are all at relatively low levels of intention and include the excavator scenario (S1), the wrong national anthem scenario (S17), the school demolition scenario (S18), and the four car accidents scenario (S11–S14).



Scenarios

S1: Graveyard
S2: Tsunami effort fails
S3: Tsunami effort succeeds
S4: Tsunami effort half succeeds
S5: Ad with kiss
S6: Ad with play on "riding"
S7: Ad with clothes addiction
S8: Record label plagiarism
S9: God comedy sketch
S10: Performance art piece
S11: Sunny day accident with person
S12: Sunny day accident with dog
S13: Windy day accident with person
S14: Windy day accident with dog
S15: Cleaner with national flag
S16: Anthem interruption
S17: Wrong anthem
S18: Demolished school flag
S19: Hispanic HR discrimination
S20: African American HR discrimination
S21: Asian HR discrimination
S22: Hispanic HR fairness
S23: African American HR fairness
S24: Asian HR fairness
S25: Hispanic college discrimination
S26: African American college discrimination
S27: Asian college discrimination
S28: Hispanic college fairness
S29: African American college fairness
S30: Asian college fairness
S31: Hispanic salary discrimination
S32: African American salary discrimination
S33: Asian salary discrimination
S34: Hispanic salary fairness
S35: African American salary fairness
S36: Asian salary fairness
S37: Hispanic police discrimination
S38: African American police discrimination.
S39: Asian police discrimination
S40: Hispanic police fairness
S41: African American police fairness
S42: Asian police fairness
A1: Fire effort success
A2: Fire effort half success
A3: Fire effort fail
A4: Hurricane effort success
A5: Hurricane effort half success
A6: Hurricane effort fail
A7: Supermarket robbery
A8: Pharmacy robbery
A9: Jewelry robbery
A10: Bank robbery
A11: Ambush
A12: Raw material shortage
A13: Excess material
A14: Good prediction
A15: Counseling
A16: Hiking
A17: Company brand names
A18: Amusement parks
A19: Ray's painkillers
A20: Ben the physical therapist
A21: Twitter
A22: Gas tax
A23: Interest rates
A24: Suspected terrorist
A25: Senior assistant
A26: Civil engineer collapse
A27: Civil engineer success
A28: Shoplifting security

Scenarios beginning with an A can be found in the appendix

Figure 6.2

Judgments of human and machine actions across scenarios.

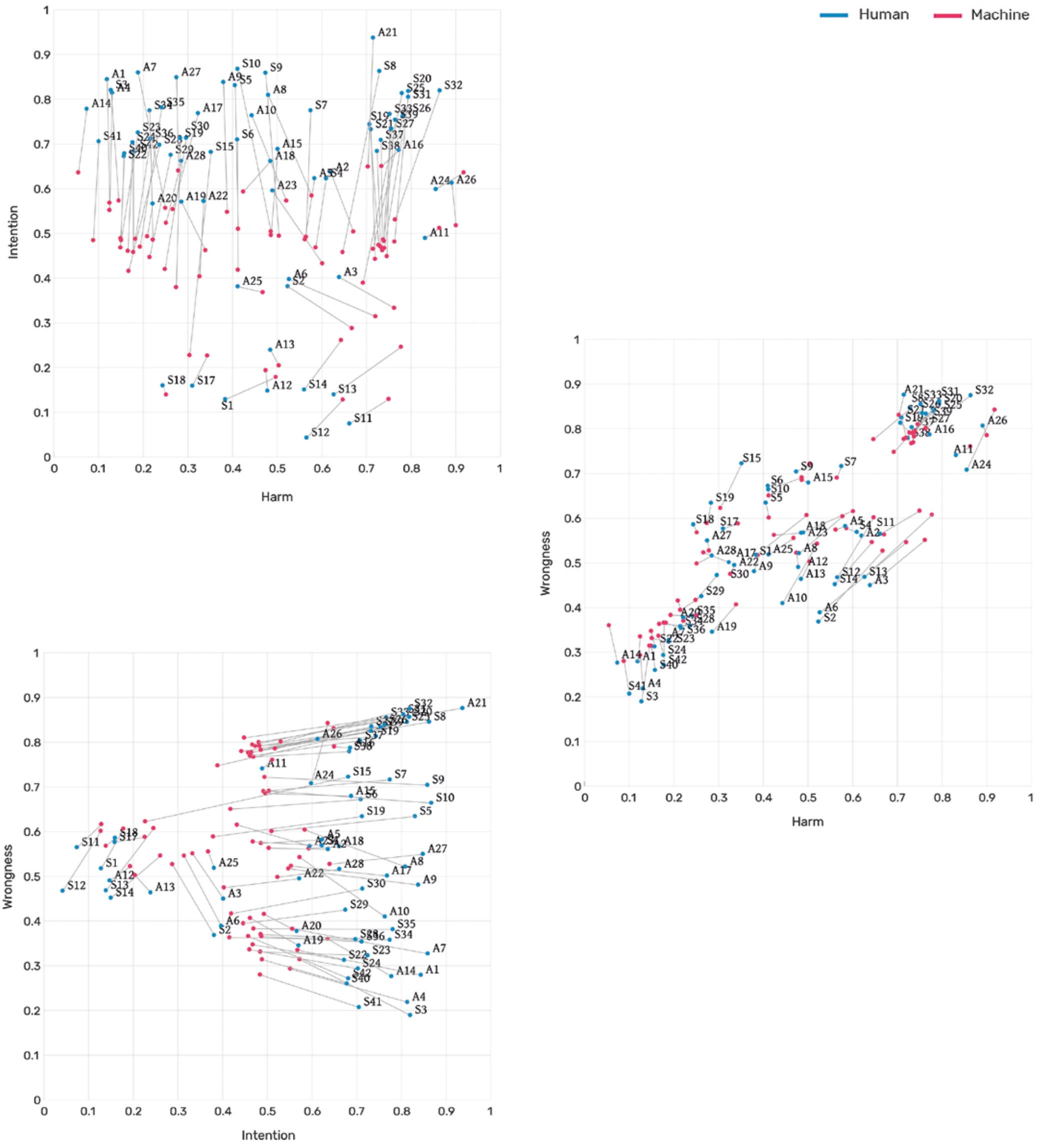


Figure 6.3

Harm-intention plane, wrongness-intention plane, and harm-wrongness plane.

The harm-intention plane reveals two things: The first, which is obvious, is that in most cases, people appear to assign more intention to human actions than machine actions. The second, which is more surprising, is that people may excuse human actions more than machine actions in accidental scenarios. For instance, when a car accident is caused by either a falling tree or a person jumping in front of a car, people assign more intention to the machine than to the human behind the wheel. As discussed in previous chapters, this suggests that people perceive an accident more like an error when the actor is a machine, but as misfortune when the actor is a human. Hence, in these types of scenarios, they forgive or excuse humans more than machines.

Figure 6.3 also shows the wrongness-intention plane. We also see a triangular pattern because intention modulates the level of perceived wrongness. Unintentional actions cluster close to the neutral value (0.5) "Neither wrong nor right." But actions perceived as intentional can score very high ("Extremely wrong") or very low ("Extremely right"). This is consistent with an extensive body of literature in moral psychology showing that intentional actions are judged worse than accidents, even when the accidents have more serious consequences.[3]

But the wrongness-intention plane also reveals some interesting patterns. For low levels of intention ($I < 0.3$), we see a clear upward slope, meaning that machine actions are perceived as both more wrong and more intentional than those of humans. This group contains the four car accident scenarios (S11–S14).

At an intermediate level of intention ($0.3 < I < 0.4$), we find actions that are perceived as less intentional for machine, but also worse. These examples include those of unlucky decisions under uncertainty, like the tsunami scenario (S2), or cases with equivalent outcomes for the fire and hurricane framings (A1 and A4).

At high levels of intention, however, differences in the intention attributed to humans and machines correlate with differences in the level of perceived wrongness. For high wrongness (> 0.75), human actions are judged as more intentional and more morally extreme (worse). This group consists of cases involving discriminatory treatment in school admissions and human resources (S19–S21, S25–S27, S31–S33, S37–S39). For low wrongness (< 0.4), machine actions are seen as less intentional, but still are judged worse than the equivalent action performed by a human. This group includes cases such as those involved in correcting unfair treatment in school admissions and human resources (S19–S21, S22–S24, S25–S27, S34–S36, S40–S42). In other words, because human actions are seen as more intentional, humans are perceived as more morally right than machines in scenarios with strong positive outcomes, and as more morally wrong than machines in scenarios with strong negative outcomes.

We look at the harm-wrongness plane (figure 6.3). Unsurprisingly, we see a strong positive correlation between perceived harm and perceived wrongness. Yet we also observe regions characterized by different regimes. For positive outcomes ($W < 0.35$), we find no big difference between the harm attributed to a machine or a human action, but we do find that machine actions are judged worse. At intermediate levels of harm and wrongness ($0.4 < H < 0.75$ and $W < 0.65$), we find actions that are perceived as more harmful and worse when performed by machines than humans. In fact, the evaluation of these scenarios is so extreme that humans are—on average—perceived to be morally right ($W < 0.5$) in situations in which machines are perceived—on average—as morally wrong ($W > 0.5$). In this region, machines are also perceived as more harmful. This cluster is populated by accidental scenarios, including the car scenarios (S11–S14), the interest rate scenario (A23 in the appendix), and the unlucky outcome of the tsunami scenario (S2). In these uncertain cases, people are less forgiving of machines and judge actions as more harmful and morally worse when they are performed by machines.

Finally, for scenarios rated high on harm and wrongness (W and $H > 0.7$), we find two groups. The first one involves cases of algorithmic bias (chapter 3), which relates to the fairness dimension of moral psychology. Here, human actions are seen as both slightly more harmful and also worse than the equivalent actions performed by a machine. The second group, which exhibits the opposite trend, consists of two cases of accidental manslaughter, such as the terrorist scenario (A24) and the ambush scenario

(A11). Here, machine actions are seen as more morally wrong than those of humans, suggesting once again that the bias against machines is modulated by a scenario's moral dimensions.

The moral space tells us that the way in which people judge the actions of machines compared to those of humans varies across scenarios. When intention and harm are low, people appear to be less forgiving of machines, evaluating their actions as worse. When intention and harm are high, however, people tend to judge human actions as worse than the equivalent machine actions.

Of course, the results presented here should be taken with a grain of salt. Despite the apparent clarity of these trends, the moral space should include factors beyond a scenario's perceived level of harm and intention. For instance, in scenarios involving a dimension of fairness, such as the algorithmic bias scenarios (chapter 3), humans are judged more harshly than machines when they do wrong and more positively when they do right. In the scenarios involving physical harm, such as the car accident (S11–S14), tsunami (S2), and manslaughter scenarios (A11 and A24), machines are judged more harshly.

Also, our list of scenarios is far from exhaustive, so there is much to be learned from additional cases. Nevertheless, these findings help us understand broad trends and differences in the way in which humans judge the actions of machines compared to the actions of other humans. But can we formally model these patterns? In the next section, we model these moral surfaces mathematically to understand more systematically when people have biases for or against machines.

Moral Surfaces

Next, we construct a statistical model that maps a scenario's level of wrongness to a level of perceived intention and harm. Our goal is to study differences in the functions mapping harm and intention to wrongness for comparable human and machine actions.

To keep things simple, we will use some very rough assumptions. Even though wrongness, harm, and intention are all affected by the treatment (i.e., they change depending on whether the scenario was an action of a human or a machine), we will use these variables together in a model. This model will estimate the level of perceived wrongness of a scenario as a function of that scenario's level of perceived intention and harm. Because the dependent and independent variables are affected by the treatment, in statistics this would be considered a *heroic* assumption—an assumption that even those using it would consider untrue. Yet we find that despite this heroic assumption, our model captures some qualitatively interesting patterns—namely, that differences in people's judgment of human and machine actions are not simple preferences for humans over machines, but involve differences in the functional forms involved. These differences are expressed in the intercept, slope, and curvature of the derived moral functions.

We use individual-level data including more than 27,000 individual responses. Our goal is to estimate the following two functions to predict the wrongness of the actions performed by humans and machines:[‡]

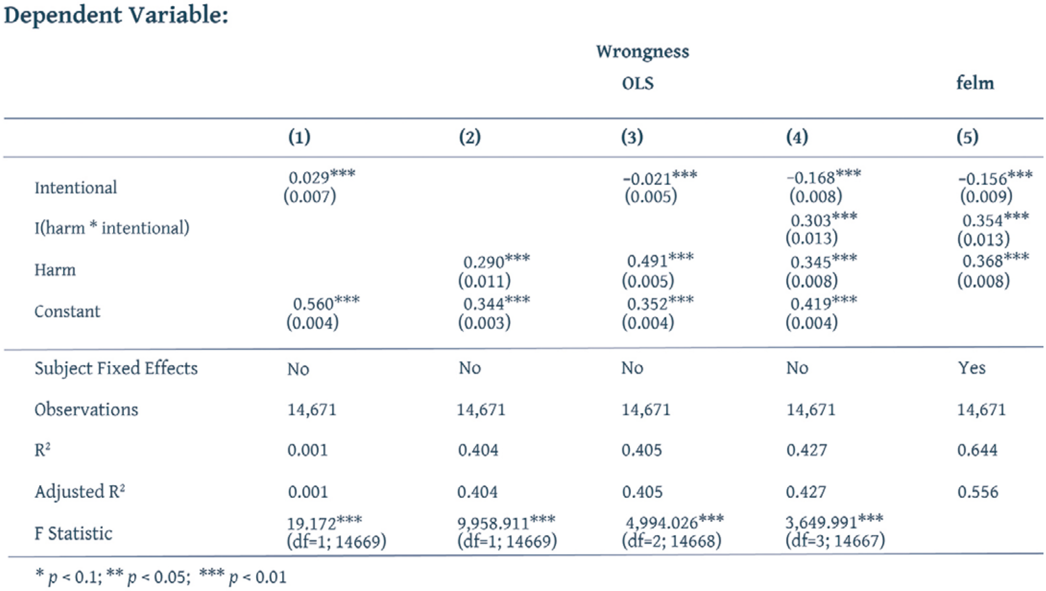
$$W = f_h(I,H)$$
$$W = f_m(I,H)$$

[‡] We could include h and m in the same function [e.g., $f(I,H,C)$, where C is the condition], but because we will be plotting the functions separately, we believe that the presentation will be clearer if we separate these functions from the beginning.

Here, the subscript *h* represents humans, and *m* stands for machines. For simplicity, we use a linear model with interactions and individual fixed effects. Using a Taylor expansion of the previous two equations, we get the following model for wrongness *W*:

$$W = B_1H + B_2I + B_3HI + \eta + \epsilon,$$

where *H* and *I* represent perceived harm and intention, η represents individual fixed effects, and ϵ is the residual. Our model includes individual *fixed effects* to capture any source of constant variation between individuals. This is a collection of vectors that are 1 for each individual and 0 for everyone else. These vectors can capture any constant source of variation among experimental subjects, such as differences in age,



Dependent Variable:

			Wrongness		
			OLS		felm
	(1)	(2)	(3)	(4)	(5)
Intentional	0.029*** (0.007)		-0.021*** (0.005)	-0.168*** (0.008)	-0.156*** (0.009)
I(harm * intentional)				0.303*** (0.013)	0.354*** (0.013)
Harm		0.290*** (0.011)	0.491*** (0.005)	0.345*** (0.008)	0.368*** (0.008)
Constant	0.560*** (0.004)	0.344*** (0.003)	0.352*** (0.004)	0.419*** (0.004)	
Subject Fixed Effects	No	No	No	No	Yes
Observations	14,671	14,671	14,671	14,671	14,671
R^2	0.001	0.404	0.405	0.427	0.644
Adjusted R^2	0.001	0.404	0.405	0.427	0.556
F Statistic	19.172*** (df=1; 14669)	9,958.911*** (df=1; 14669)	4,994.026*** (df=2; 14668)	3,649.991*** (df=3; 14667)	

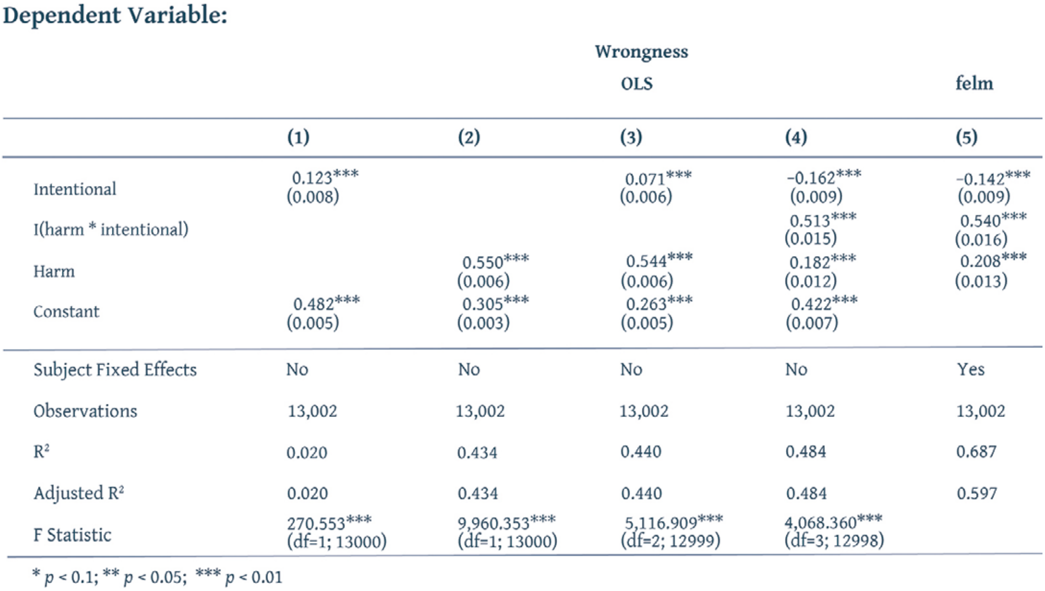
* $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$

Table 6.1

Moral functions of people judging machine actions.

gender (nonbinary), languages spoken, race, or even shoe size. Fixed effects also help us consider variations in the level of judgment of individuals, such as some individuals being too "judgy," and rating all actions too harshly, or individuals being too lenient and judging everything lightly.

Tables 6.1 and 6.2 present, respectively, the results of the models for judging machine and human actions. We introduce each term sequentially to study how the coefficients change as we move from a bivariate model (including only harm or intention) to a model with interactions and fixed effects. We find empirically that quadratic terms do not improve the predictive power of the model enough to be considered, so we drop them from the regression.



Dependent Variable:

			Wrongness		
			OLS		felm
	(1)	(2)	(3)	(4)	(5)
Intentional	0.123*** (0.008)		0.071*** (0.006)	−0.162*** (0.009)	−0.142*** (0.009)
I(harm * intentional)				0.513*** (0.015)	0.540*** (0.016)
Harm		0.550*** (0.006)	0.544*** (0.006)	0.182*** (0.012)	0.208*** (0.013)
Constant	0.482*** (0.005)	0.305*** (0.003)	0.263*** (0.005)	0.422*** (0.007)	
Subject Fixed Effects	No	No	No	No	Yes
Observations	13,002	13,002	13,002	13,002	13,002
R^2	0.020	0.434	0.440	0.484	0.687
Adjusted R^2	0.020	0.434	0.440	0.484	0.597
F Statistic	270.553*** (df=1; 13000)	9,960.353*** (df=1; 13000)	5,116.909*** (df=2; 12999)	4,068.360*** (df=3; 12998)	

* $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$

Table 6.2

Moral functions of people judging human actions.

The first four columns of these tables show the results of ordinary least squares (OLS) models. The last column shows the results of the fixed effects models (felm), which account for differences in individual characteristics.

The first two columns show the coefficients for models that include only intention and harm. The models considering only intention have no predictive power ($R^2 \leq$ 2 percent), while the models using harm as a predictor already explain a considerable amount of variance for both machine and human actions ($R^2 >$ 40 percent). Models 3 and 4 use both intention and harm, and model 4 also includes an interaction term for harm and intention. Adding the interaction term increases the amount of variance explained by the models to 43 percent in the machine scenarios and 48 percent in the human scenarios. Finally, the felm models explain 56 percent of the variance in the machine condition and 60 percent in the human condition (adjusted R^2).

Even though the fixed effects model explains significantly more variance than the OLS, the coefficients associated with harm, intention, and their interaction do not vary drastically.[§] This means that the coefficients of the model are not greatly biased by differences in individual characteristics.

To interpret these coefficients, we visualize the planes defined by the fourth column of each table (figures 6.4, 6.5, and 6.6), as well as the cross sections (figures 6.7 and 6.8). We find that the hyperplanes respect some of the characteristics observed in the moral space, and hence serve as crude empirical models of moral functions.

[§] The harm coefficient (B_1) changes from 0.345 or 0.368 for machines, and from 0.182 and 0.208 for humans. The intention coefficients (B_2) are –0.168 and –0.156 for machines and –0.163 and –0.142 for humans. The interaction coefficients are 0.303 and 0.354 for machines and 0.513 and 0.540 for humans.

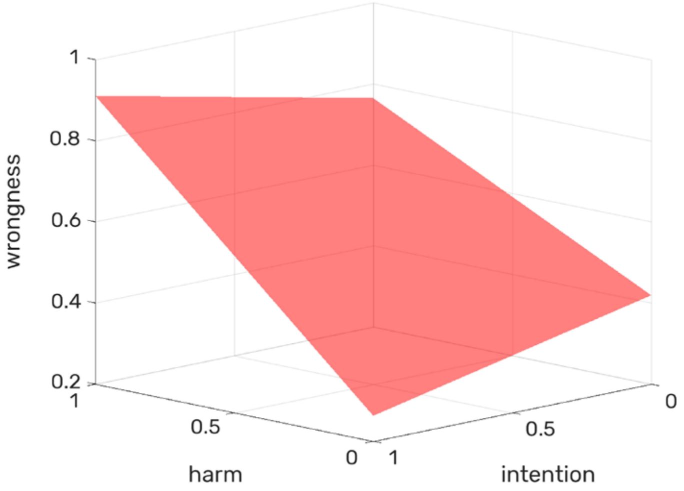


Figure 6.4

Moral functions of people judging machine actions.

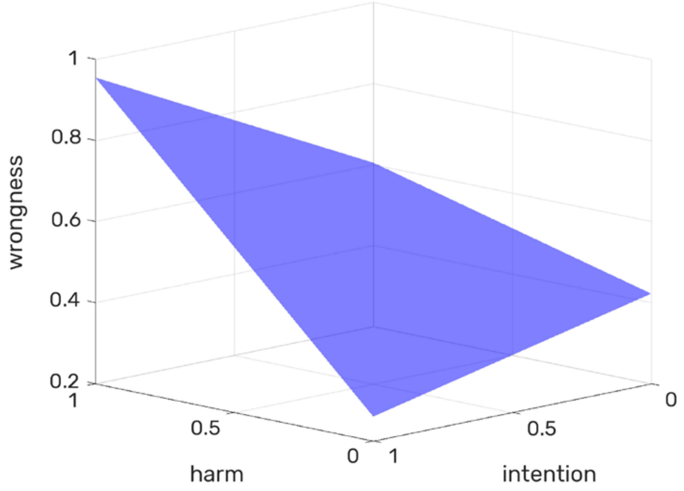


Figure 6.5

Moral functions of people judging human actions.

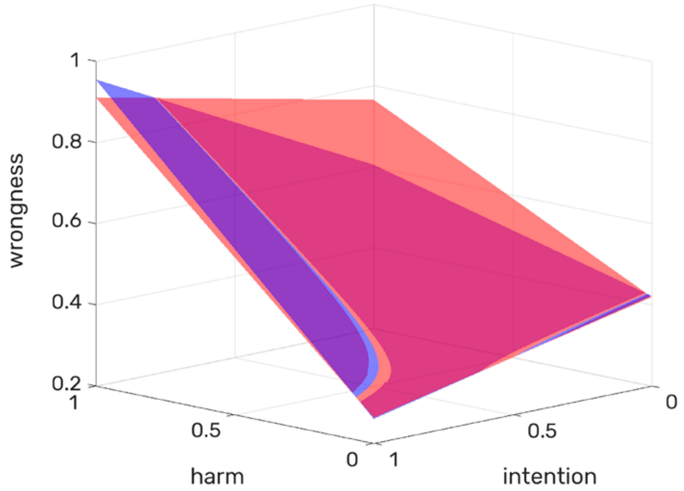


Figure 6.6

Visualization of the moral functions described in tables 6.2 and 6.3.

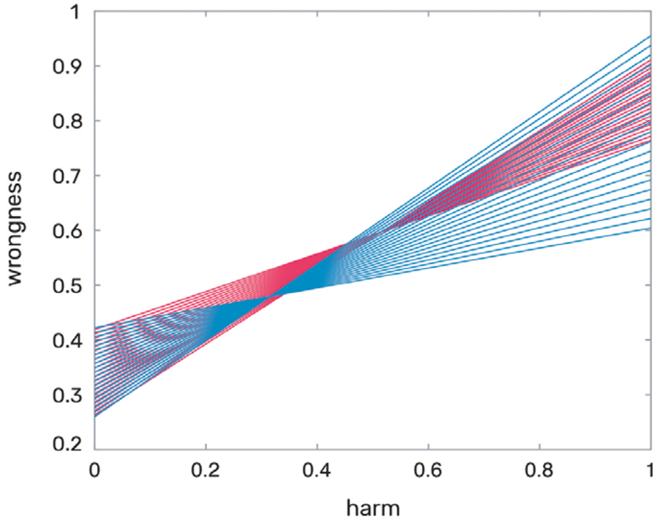


Figure 6.7

Cross section of moral functions in the wrongness and harm planes.

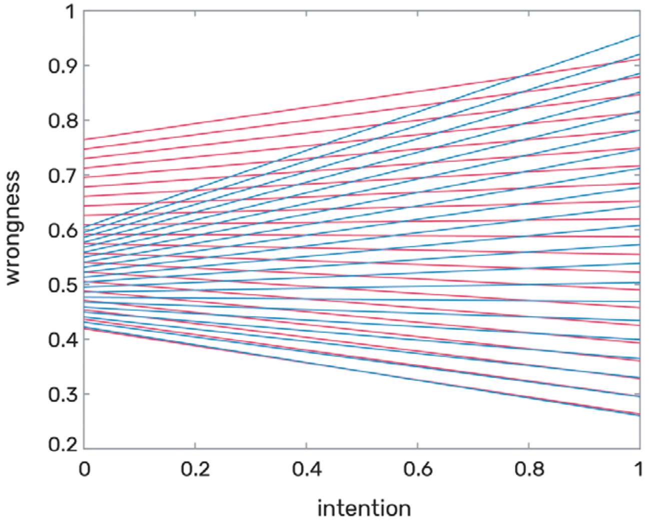


Figure 6.8

Cross section of moral functions in the wrongness and intention planes.

Figure 6.8 shows that intention enhances the perceived wrongness of human actions more than that of machines. This comes mostly from the interaction term (harm × intention). For machines, the slope of wrongness on harm is the dominant feature of the model, suggesting that **humans are judged by their intentions, while machines are judged by their outcomes.** Of course, this is a simplification, since the interaction between intention and harm is also significant in the model of humans judging machines. But to a first approximation, these differences in the relative importance of coefficients describe, coarsely and qualitatively, the difference between these two moral functions.

Also, we find that at high levels of harm and intention, human actions are judged more harshly. This is observed in the fanning out of the wrongness-intention curves for different levels of harm (figure 6.7). As a result of that, humans appear to judge the actions of other humans more harshly at the highest levels of harm and intention, but they judge machines more harshly in the rest of this space. Certainly, this is not applicable to all cases—it is a crude approximation—but it is an aggregate description that can serve as a quick rule of thumb to think about differences in human and machine judgment.

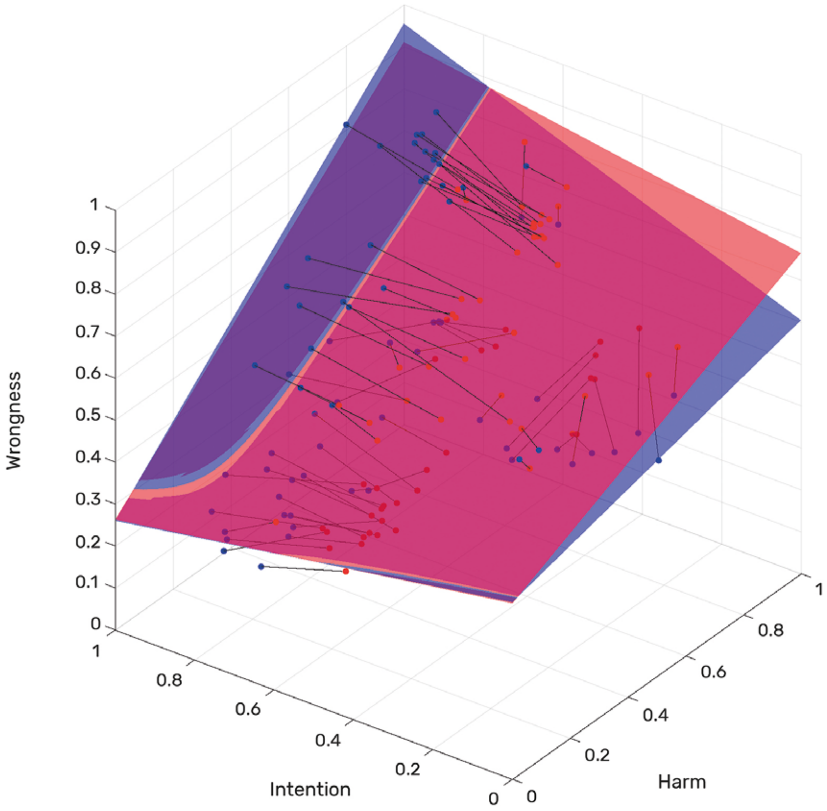


Figure 6.9

Model compared to empirically observed means.

Finally, we compare this model—trained with individual data—to the empirically observed means (figure 6.9). The model appears to capture a good deal of the variance observed in the moral judgment of scenarios and, more important, it also tends to capture the direction of the treatment effect. Yet, because this is a regression model, the empirical values tend to be over or under the estimated hyperplane (regression to the mean), meaning that the model underestimates the wrongness of the worst scenarios or the goodness of the best ones.

But are these judgments affected by the characteristics of the observers? Do people with different ethnicities, genders, or levels of education judge things differently? Are some of these groups more inclined to judge machines or humans more harshly? In the next section, we continue our statistical exploration by looking instead at how the demographics of experimental subjects correlate with their judgments of humans and machines.

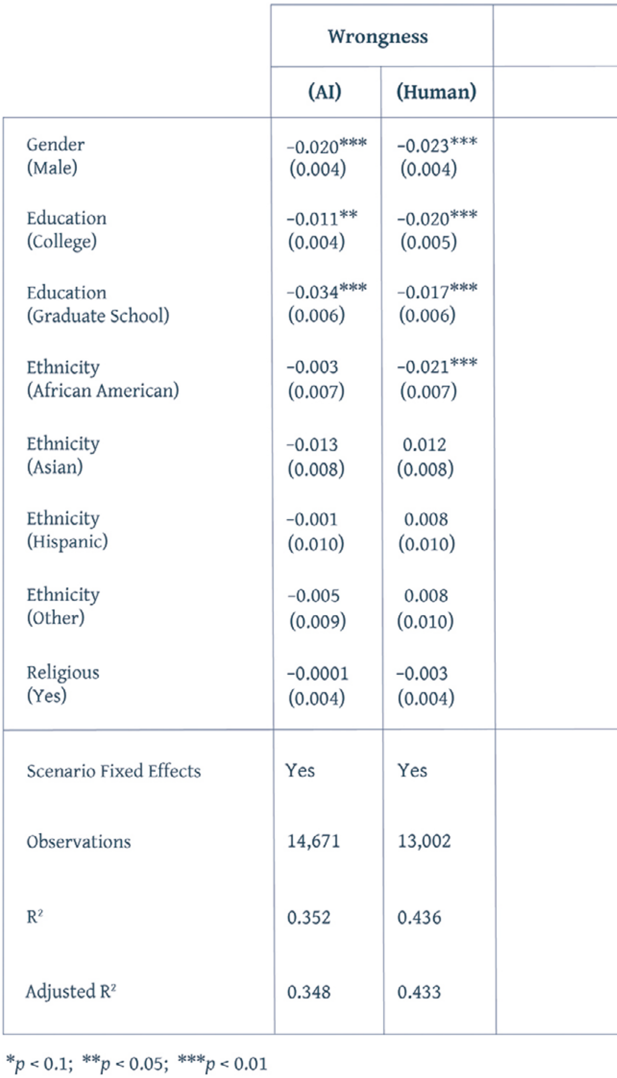
Who Is the Judge?

In this section, we study how different demographic characteristics, such as the gender, ethnicity, and education of subjects, correlate with their answers to the questions provided for each scenario. We focus on six questions:

- How morally **wrong** or right is the agent's action/decision?
- How **harmful** is the action/decision?
- How **intentional** is the action/decision?
- How much do you **like** the agent?
- If you were in a **similar situation**, would you have done the same?
- Do you agree that this (person/machine) agent should be replaced (machine/person)? (**replaced different**)

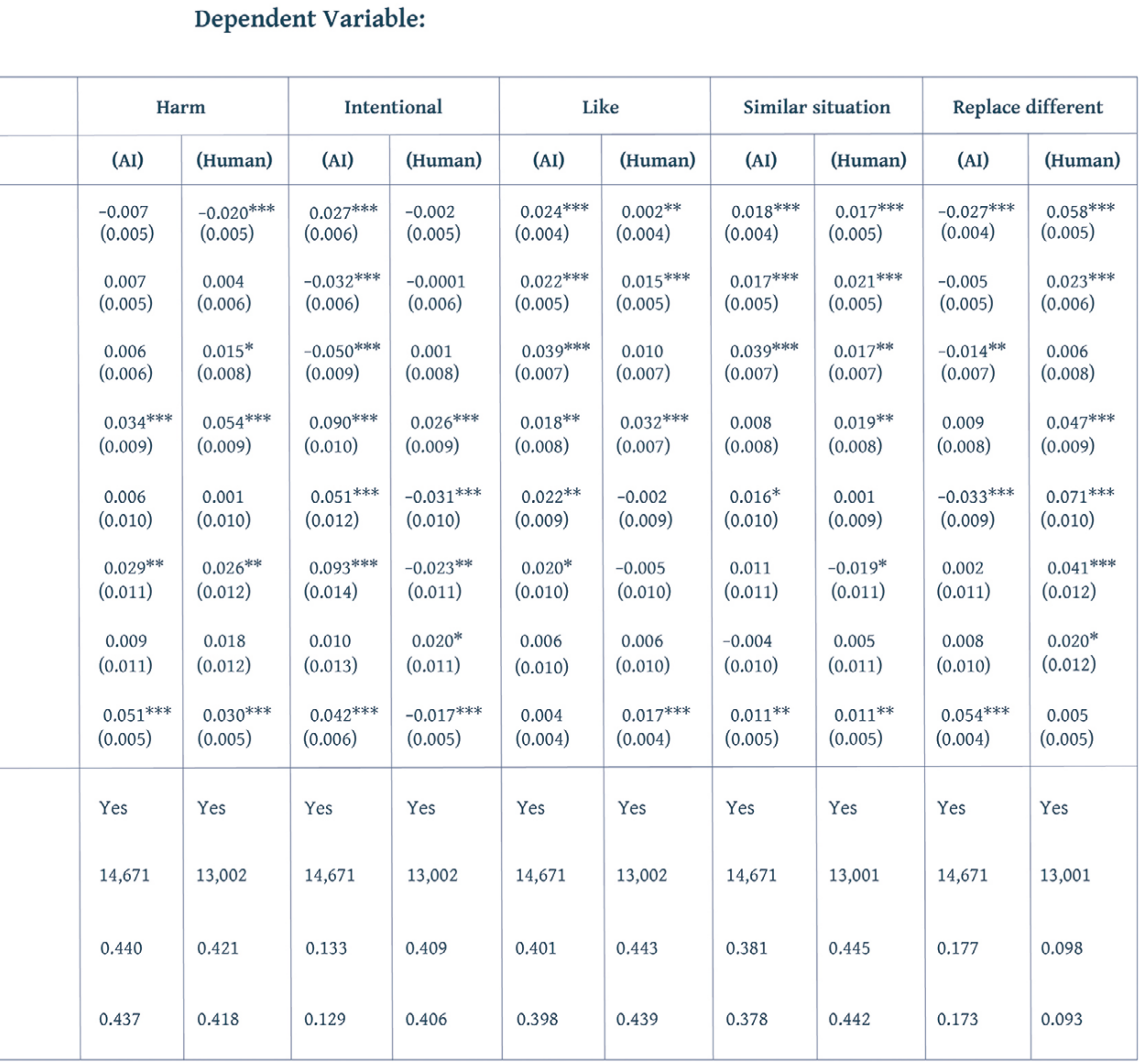
We explore how the answers to these questions correlate with the demographic characteristics of individuals. To do this, we construct a model with scenarios as fixed effects. Scenario fixed effects models include vectors that are 1 for each scenario and 0 for all others. These vectors capture any constant variations between scenarios (such as the average response received by each of them). After controlling for scenario fixed effects, the variables on the demographic dimensions should capture variations in judgment that are not explained by the scenario itself, but rather by the characteristics of the respondents.

We looked at four individual characteristics: people's **gender** (using a nonbinary description of male, female, and other), level of **education** (high school, college, and graduate school), **ethnicity** (white, African American, Asian, Hispanic, and other), and whether people self-report as **religious** (yes or no). Because of data sparsity, we considered only "Male" and "Female" answers for gender (only two survey respondents answered "Other").



	Wrongness	
	(AI)	**(Human)**
Gender (Male)	−0.020*** (0.004)	−0.023*** (0.004)
Education (College)	−0.011** (0.004)	−0.020*** (0.005)
Education (Graduate School)	−0.034*** (0.006)	−0.017*** (0.006)
Ethnicity (African American)	−0.003 (0.007)	−0.021*** (0.007)
Ethnicity (Asian)	−0.013 (0.008)	0.012 (0.008)
Ethnicity (Hispanic)	−0.001 (0.010)	0.008 (0.010)
Ethnicity (Other)	−0.005 (0.009)	0.008 (0.010)
Religious (Yes)	−0.0001 (0.004)	−0.003 (0.004)
Scenario Fixed Effects	Yes	Yes
Observations	14,671	13,002
R^2	0.352	0.436
Adjusted R^2	0.348	0.433

*$p < 0.1$; **$p < 0.05$; ***$p < 0.01$



Dependent Variable:

	Harm		Intentional		Like		Similar situation		Replace different	
	(AI)	(Human)	(AI)	(Human)	(AI)	(Human)	(AI)	(Human)	(AI)	(Human)
	-0.007	-0.020***	0.027***	-0.002	0.024***	0.002**	0.018***	0.017***	-0.027***	0.058***
	(0.005)	(0.005)	(0.006)	(0.005)	(0.004)	(0.004)	(0.004)	(0.005)	(0.004)	(0.005)
	0.007	0.004	-0.032***	-0.0001	0.022***	0.015***	0.017***	0.021***	-0.005	0.023***
	(0.005)	(0.006)	(0.006)	(0.006)	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)	(0.006)
	0.006	0.015*	-0.050***	0.001	0.039***	0.010	0.039***	0.017**	-0.014**	0.006
	(0.006)	(0.008)	(0.009)	(0.008)	(0.007)	(0.007)	(0.007)	(0.007)	(0.007)	(0.008)
	0.034***	0.054***	0.090***	0.026***	0.018**	0.032***	0.008	0.019**	0.009	0.047***
	(0.009)	(0.009)	(0.010)	(0.009)	(0.008)	(0.007)	(0.008)	(0.008)	(0.008)	(0.009)
	0.006	0.001	0.051***	-0.031***	0.022**	-0.002	0.016*	0.001	-0.033***	0.071***
	(0.010)	(0.010)	(0.012)	(0.010)	(0.009)	(0.009)	(0.010)	(0.009)	(0.009)	(0.010)
	0.029**	0.026**	0.093***	-0.023**	0.020*	-0.005	0.011	-0.019*	0.002	0.041***
	(0.011)	(0.012)	(0.014)	(0.011)	(0.010)	(0.010)	(0.011)	(0.011)	(0.011)	(0.012)
	0.009	0.018	0.010	0.020*	0.006	0.006	-0.004	0.005	0.008	0.020*
	(0.011)	(0.012)	(0.013)	(0.011)	(0.010)	(0.010)	(0.010)	(0.011)	(0.010)	(0.012)
	0.051***	0.030***	0.042***	-0.017***	0.004	0.017***	0.011**	0.011**	0.054***	0.005
	(0.005)	(0.005)	(0.006)	(0.005)	(0.004)	(0.004)	(0.005)	(0.005)	(0.004)	(0.005)
	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
	14,671	13,002	14,671	13,002	14,671	13,002	14,671	13,001	14,671	13,001
	0.440	0.421	0.133	0.409	0.401	0.443	0.381	0.445	0.177	0.098
	0.437	0.418	0.129	0.406	0.398	0.439	0.378	0.442	0.173	0.093

Table 6.3

Model coefficients for demographic characteristics.

Because these are all categorical variables, we measured their effects using a reference level. For gender, we show the coefficients of the Male category in reference to the Female category (i.e., only Male shows up in the regression results because the coefficient reports a difference between the two categories). In the case of education, we compare the responses of subjects with college and graduate school education relative to those with high a school education. In the case of ethnicity, we use white as a baseline.

Table 6.3 and figure 6.10 show the results of these statistical models. The odd columns (1, 3, and so on) have coefficients for the machine condition, and the even columns (2, 4, and so on) have coefficients for the human condition. These coefficients represent how much that variable increases or decreases judgment in a dimension (e.g., harm and like) after controlling for each scenario's characteristics.

One variable that does correlate with some judgments is gender. Compared to females, males tend to rate both machine and human scenarios as less morally wrong and are more likely to report having done the same in a "similar situation." Where the effects of gender appear stronger, however, is in the "replace by different" dimension, which is the question that asks people if they would replace a machine by a human or a human by a machine. Our data reveal that males are more prone to replace humans by machines and less prone to replace machines by humans.

Another variable that shows strong correlations is education. People with a college or graduate degree see the human and machine scenarios as less morally wrong than people with a high school education. This effect is particularly strong for people with a graduate degree judging machine actions. People with a college or graduate degree also see machine actions as less intentional than high school graduates and report liking machines and humans more. People with college and graduate degrees also think of themselves as more likely to have done the same action in a similar situation.

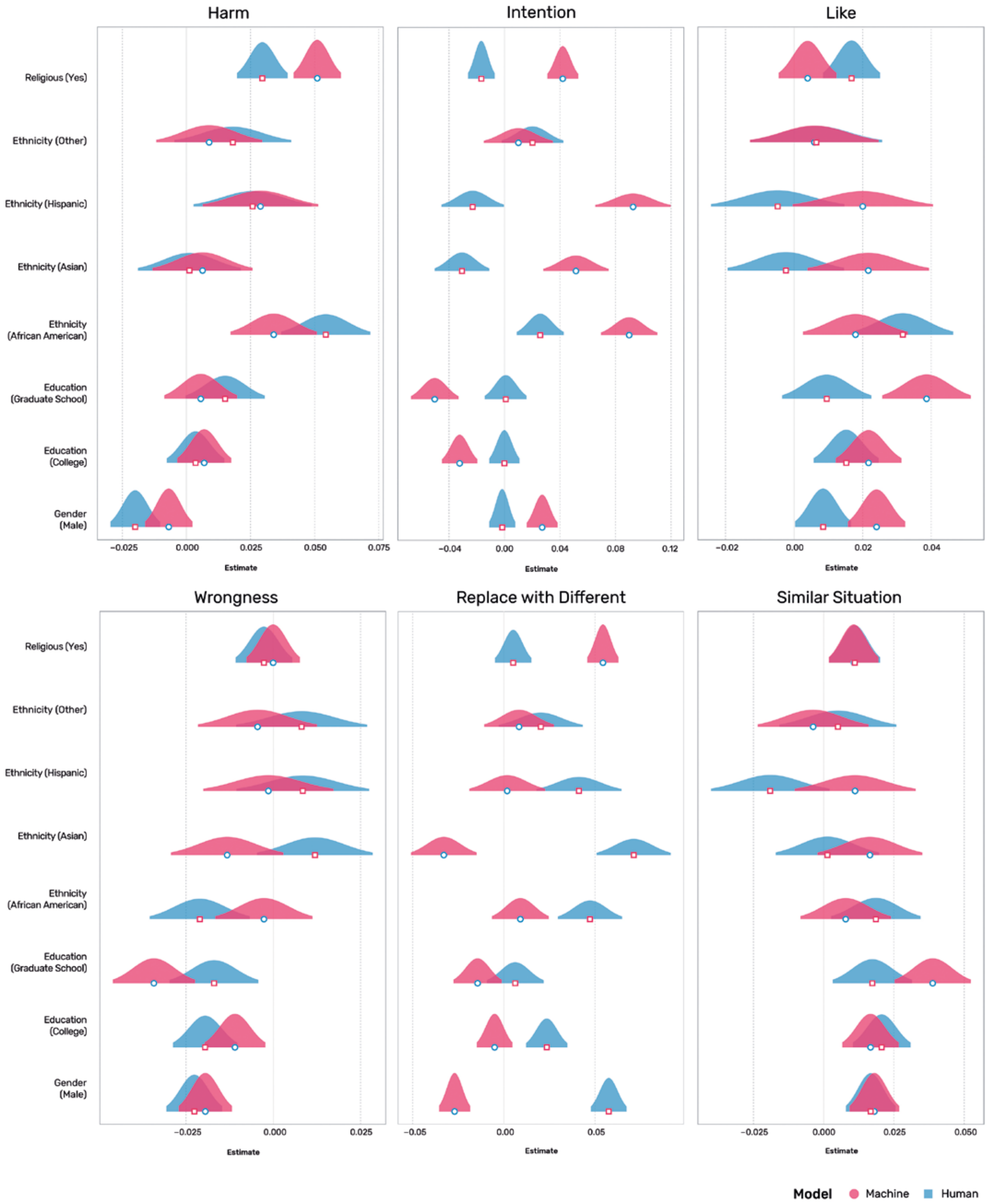


Figure 6.10

Demographic effects on the judgments of human and machine actions: harm, intention, like, moral, replace with different, and similar situation.

When it comes to ethnicity, we find differences, especially in the intention dimension. African Americans, Asians, and Hispanics attribute more intention to machine actions. Asians also show a strong effect in the "replace by different" question, showing a preference in favor of machines.

Together, these findings tell us that—on average—demographic characteristics correlate with judgments. Yet, the effects of demographics are relatively weak, shifting judgments by about 0.05 in variables that range from 0 to 1. This is consistent with the finding that individual fixed effects do not change drastically the coefficients of moral functions. Still, together, these effects can compound to create noticeable differences. For instance, a religious Hispanic male would—on average—assign 0.16 more intention to a machine than a nonreligious white female.

Discussion

In this chapter, we abstracted away from individual scenarios to provide a statistical description of the patterns that emerge across them. This exploration was split into three sections.

First, we introduced the moral space to conduct a descriptive exercise that looked at each scenario using data on harm, intention, and wrongness. It helped us confirm some observations that had emerged when discussing some scenarios. For instance, the exercise showed that humans judge the intentions of other humans using a bimodal distribution, but judge the intention of machines using a unimodal distribution. This means that people are more willing to forgive humans for accidental situations, but also attribute intent to human actions that cannot be easily excused as accidental. This is particularly true in scenarios focused on fairness, like those presented in chapter 3. We also found that people judge machine actions harshly (in terms of both harm and wrongness) in scenarios involving accidents that lead to physical harm (e.g., the self-driving car and tsunami scenarios), suggesting that people judge machines based on outcomes and judge humans based on intentions.

Our second and third exercise used fixed effects models. The second exercise used fixed effects for participants to model the relationship between a scenario's wrongness and its perceived level of intention and harm. The third exercise explored how judgments vary based on the demographic characteristics of the study's participants.

The second exercise helped us formalize some of the patterns observed in our descriptive analysis. We found different moral functions describing people's judgments of machine and human actions. Overall, people tend to judge machines more harshly across most of this space, except for scenarios with high levels of intention and harm. In fact, the main difference between the functions describing judgments of human and machine actions is whether harm, or the interaction between harm and intention, carries more weight in the model. For machines, harm tends to be the most important predictor of moral judgment. For humans, the most important predictor is the interaction term between intention and harm.

The third exercise taught us that judgments vary with demographic characteristics, although these variations are relatively mild.

Once again, these findings suggest that people judge machines based on the observed outcome, but judge humans based on a combination of outcome and intention.

In the next chapter, we conclude our journey by drawing some lessons from works of fiction and summarizing some of our main findings. This will conclude our exploration of how humans judge machines.