A Rule Based Model of Judging Harm-doing[1]

Thomas R. Shultz

McGill University

ABSTRACT

A rule based computational model of the judgment of harm-doing is presented that qualitatively simulates the major principles of an emerging psychological theory of common sense moral reasoning. Simulation results indicate that the model, called MR for Moral Reasoner, generates verdicts in substantial agreement with those reached in somewhat difficult court cases. A higher rate of agreement with outcomes produced in simpler cases from traditional cultures suggests that the model possesses a good deal of cultural universality. Systematic damaging of the rules in the model indicated that most of the rules are essential in producing a high rate of agreement with court decisions and identified some rules regarding the mental state of the accused that, individually, are less essential because they compensate for each other.

A PSYCHOLGICAL THEORY OF JUDGING HARM-DOING

This project concerns the common sense evaluation of harm-doing. Whenever a person may have been harmed by someone else, a number of issues naturally arise. How was the harm caused, is anyone responsible for the harm, is that person blameworthy, and how much should he be punished? People encounter such cases of harm-doing frequently, either directly or through secondary accounts.

Shultz and Schleifer (1983) have developed a psychological theory of reasoning about harm-doing that was inspired principally by conceptual analyses in jurisprudence and moral philosophy. A brief synopsis of this theory is presented here, minus the philosophical motivation and supporting psychological evidence, much of which is reviewed in Darley and Shultz (1990). The main concepts and decisions in the theory are illustrated in Figure 1. For a case in which a person may have done something to harm someone else, major decisions focus on causation, morally responsibility, blame, and punishment. Each of these major decisions presupposes and uses information from previous major decisions.

Judgments of moral responsibility, for example, presuppose those of causation. If the accused is judged not to have caused the harm, then there is no need to consider whether he is morally responsible for it. Similarly, judgments of blame presuppose those of moral responsibility, and decisions about punishment presuppose those about blame. A person is responsible for harm that he caused if the harm cannot be excused. Blame refers to a decision that a person is at fault, given that he has caused and is responsible for the harm. Without responsibility, there is no need to consider blame. Punishment refers to a decision about what consequences should befall the person as a result of being blameworthy. If the person is blameless, then no decision needs to be taken about punishment.

---

[1] I am grateful to John Darley and Kevin Dunbar for helpful comments on this research. This research is supported by a grant from the Social Sciences and Humanities Research Council of Canada. Address correspondence to Thomas R. Shultz, Department of Psychology, McGill University, 1205 Penfield Avenue, Montréal, Québec, Canada H3A 1B1. E-mail: ints@musicb.mcgill.ca

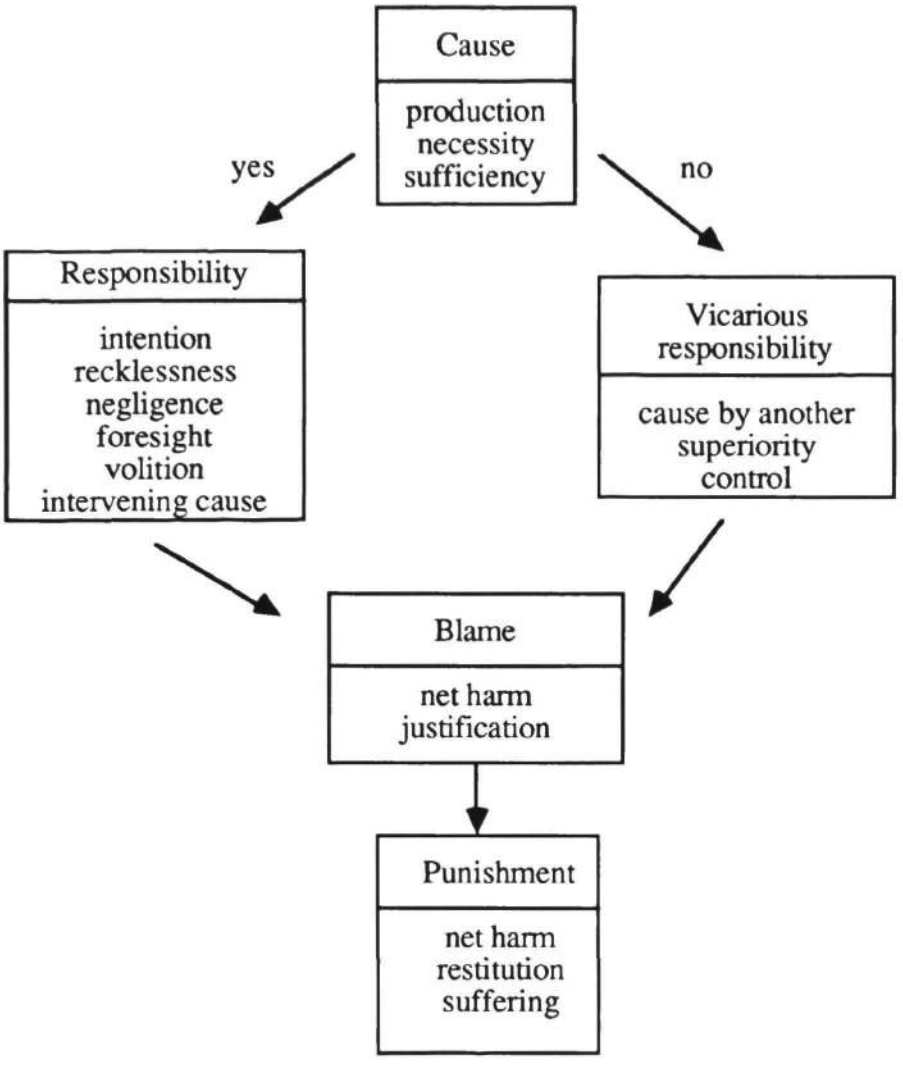

Figure 1. Major decisions and concepts in a theory of judging harm-doing

The first major judgment to be made concerns the causation of the harm. Causation is determined by a combination of generative and conditional information. On the generative view of causation, an effect is considered to be produced by some sort of transmission from the cause. Supplementing this approach is the *but for* (or sine qua non) test, which is widely used in jurisprudence. The *but for* test holds that a person's behavior is a cause of harm if and only if the harm would not have occurred without the person's behavior, thus focusing on the necessary conditions for harm. Some have argued for the use of sufficient conditions in judging causation: does the person's action distinguish the current harm-producing situation from some appropriate standard in which harm did not result?

The next major decision, responsibility, is determined by joint consideration of causation and excuses. One is held morally responsible for harm that he caused unless the harm was done accidentally (i.e., without intention, recklessness, or negligence), involuntarily (i.e., under external force), or without being able to foresee the resulting harm. Moreover, the causal chain leading from the action to the harm cannot be broken by some unforeseen event which exacerbated the harm.

The preferred way of judging intention is to match the accused's plan against the outcome. If his plan is known and the harm is included in the plan, then it is concluded that the harm was intended unless the harm was not caused as planned. If the actor's plan is unknown, then one falls back to objective heuristics such as valence, monitoring, and discounting. Did the harmful outcome have positive consequences for the actor, did he monitor his actions, or can intentionality be discounted because of alternate external causes? Recklessness is acting without due care coupled with high foreseeability of harm, unless the harm was intended. Negligence is acting without due care, but with a lower foreseeability of harm, unless the harm was intended or done recklessly.

Blame is a joint function of moral responsibility, the presence of some degree of net harm, and justification for the harm. If the accused is morally responsible for the harm and there is some net harm (i.e., more harm than benefit to the victim) then the accused is blameworthy, unless the harm can be justified. The distinction between moral responsibility and blame is somewhat subtle, but relies on the difference between excuses and justifications. Excuses, reflecting the concepts in the responsibility box in Figure 1, are offered when one admits to having caused harm, but does not accept responsibility for it. If such an excuse is accepted, then the issue of blame does not arise. Justifications come into play when one accepts responsibility for having caused harm, but denies that it was bad thing to have done, thereby avoiding blame, and perhaps earning credit.

In order to be justified, harm must achieve some goal, that goal must be more highly valued than not doing the harm, and the goal must be achievable in no less harmful way. These conditions are evaluated in sequence since a consideration of each one presupposes an appropriate decision on the preceding one. For example, if the harm achieves no goal, then there is no reason to consider whether any goal is more highly valued than not doing the harm.

The discussion has so far focussed on holding someone blameworthy for harm that he has directly caused. However, it is possible for blame to be assigned without direct causation. Such cases are typically understood as involving vicarious responsibility. A person is held vicariously responsible only when that person is in a superior position to the perpetrator or could have prevented the perpetrator from causing harm.

If the accused is blameworthy, then punishment can be assigned. Consistent with the retribution theory of punishment, punishment is directly proportional to the net amount of harm, scaled down by restitution the perpetrator has made, and the degree to which the perpetrator has suffered as a result of having caused the harm.

COMPUTATIONAL MODEL

A computer program was developed to simulate how the ordinary person (down to about 5-years-old) reasons about harm-doing. The program is called MR, for Moral Reasoner. MR provides a convenient way of rigorously specifying the psychological theory reviewed above and a technique for having the theory generate conclusions that can be compared with those produced by human subjects.

The version of MR used here is written in Lisp with rules implemented as boolean procedures, returning either *true* or *false*.[2] An English version of the rule dealing with moral responsibility is given as an example:

---

[2] A more conventional way to write this program would be as production rules in a production system interpreter. Several versions of MR have been done in just that way.

If & the accused produced the harm
    the accused's action was not accidental
    the accused's action was voluntary
    the harm was a foreseeable consequence of the accused's action
    there was no intervening cause of the harm
Then the accused is morally responsible for the harm
Else the accused is not morally responsible for the harm

The MR program accepts a case described in terms of categorical values on a number of features (e.g., foreseeability of the harm is high) and produces a series of conclusions on any of the other critical concepts in the model needed or requested.

EXAMPLE TRACE FOR A SINGLE CASE

The following illustrates how the MR program deals with a particular case. The case of Lynch vs. Fisher was tried in Louisiana in 1947 (Hart & Honoré, 1959):

A highway collision occurred through the negligence of accused, whereby a third party was trapped in his car and injured. The plaintiff, seeing the collision, went to help and finding a pistol on the floor, handed it to the injured man, who in a state of delirium through the shock of the accident, fired at plaintiff and wounded him.

A case is described to the MR program as a set of attribute-value pairs. The particular values for Lynch vs. Fisher[3] were:

((case-name lynch v fisher) (produce-harm ?) (necessary-for-harm y) (sufficient-for-harm n) (mental-state negligent) (careful n) (plan-known n) (plan-include-harm ?) (harm-caused-as-planned ?) (monitor y) (benefit-accused n) (foreseeability low) (external-cause n) (external-force n) (intervening-contribution y) (foresee-intervention n) (severity-harm 0.5) (benefit-victim 0) (achieve-goal n) (goal-outweigh-harm ?) (goal-achieveable-less-harmful ?) (restitution 0) (accused-suffer 0) (verdict g))

The output from the MR program is presented as a series of inferences. For Lynch vs. Fisher, the output was as follows: Case of Lynch v Fisher; accused caused the harm; no direct evidence that accused intended the harm; accused's intention to harm cannot be discounted; accused was not reckless; accused was negligent; no indirect evidence that accused intended the harm; accused did not intend the harm; the harm was not accidental; accused's action was voluntary; the harm was foreseeable; there was an intervening cause of the harm; accused is not morally responsible for the harm; accused is not blameworthy; disagree.

The accused caused the accident negligently, thus ruling out a pure accident. Also, the accused's actions were voluntary (i. e., not forced) and some kind of serious harm was

---

However, the Lisp version described here was found to be especially convenient for simulating large numbers of cases with damaged rules (see DAMAGING THE MODEL, below). The Lisp version of MR functions much like a backward chaining production rule interpreter.

[3] The symbols y, n, and ? refer to yes, no, and undecided, respectively. The symbol g refers to guilty, the verdict being used only to tabulate the rate of agreement between the program and the court.

foreseeable from negligent driving. However, because there was an intervening contribution to the harm, which was not foreseeable, MR decides that there was an intervening cause of the harm that mitigates responsibility and, thus, blame. In contrast, the court found the accused guilty.

EVALUATING THE COMPUTATIONAL MODEL

The model was tested on two large sets of actual cases of harm-doing. One set was based on legal cases in English and American law; the other set on cases recorded among traditional cultures that possess no codified legal system.

The first set consisted of the 95 most fully described legal cases in Hart and Honoré (1959), spanning the last five centuries of Anglo-American law. The overall proportion of agreement on MR's decision of blameworthy with a judicial decision of guilty was .84, $X^2 = 44.47, p < .001$.[4] This rate of agreement is gratifyingly high considering the difficulty of this set of cases, as indicated by the relatively low proportion of unsuccessful appeals, .61. The model agreed more with the final judicial decision than did the initial judicial decision, $X^2 = 12.15, p < .05$.

The second set contained 58 cases of harm-doing reported by anthropologists working in traditional, non-literate cultures around the world. The largest, single source was Pospisil (1958). The proportion of agreement was higher here (.97) than with the Hart and Honoré cases, $X^2 = 4.61, p < .05$, reflecting the fact that these cases were conceptually quite simple. This result suggests that the model implemented in MR does have some claim to cultural universality.

These above simulations, initially conducted blind with regard to the real life decisions, were useful in fine tuning the rule base, chiefly by repairing inconsistencies and anomalies in the rules.

DAMAGING THE MODEL

In order to determine whether each of the rules in the model is critical to matching real-life decisions, the cases were run again under conditions in which each rule was damaged. A rule was damaged by reversing the boolean value it returned: *true* if it was supposed to return *false*, *false* if it was supposed to return *true*. Only those rules leading up to the critical decision of blameworthy were damaged in this way.

The results for the 95 Hart and Honoré cases are presented as solid bars in Figure 2 in terms of proportion agreement with the final judicial decision. The proportion agreement produced by each type of damage was contrasted with that produced by the fully intact model using log-linear analysis. Damage to each rule did significantly lower the agreement rate (mean = .35, all $ps < .001$), except for those rules dealing with the mental state of the accused (mean = .83, all $ps > .5$). This revealed that for decisions of blame in cases of non-accidental harm (virtually all of the Hart and Honoré cases), it does not matter whether the accused acts intentionally, recklessly, or negligently. This seemed quite surprising until it was realized that these mental state concepts compensate for each other so that, even if one is damaged, another fills the gap.

---

[4] All of the statistics reported in this paper are based on log-linear analysis. Each result is reported as a 1 *df* Wald chi-square value.

Two rules are critical to understanding this compensation. One rule deals with accident, the other with intention:

If the protagonist's action was intentional, reckless, or negligent
Then the harm was not accidental
Else the harm was accidental

If there is direct or indirect evidence that the protagonist intended the harm
Then the harm was intended
Else the harm was not intended

As an example of compensation, even if the *intend* rule is damaged, the *reckless* or *negligent* rule may still prevent the harm from being viewed as accidental. As another example, even if the *strong-intend* rule (providing direct evidence of intention) is damaged, the *weak-intend* rule (providing indirect evidence of intention) may still lead to the conclusion that the harm was intended.

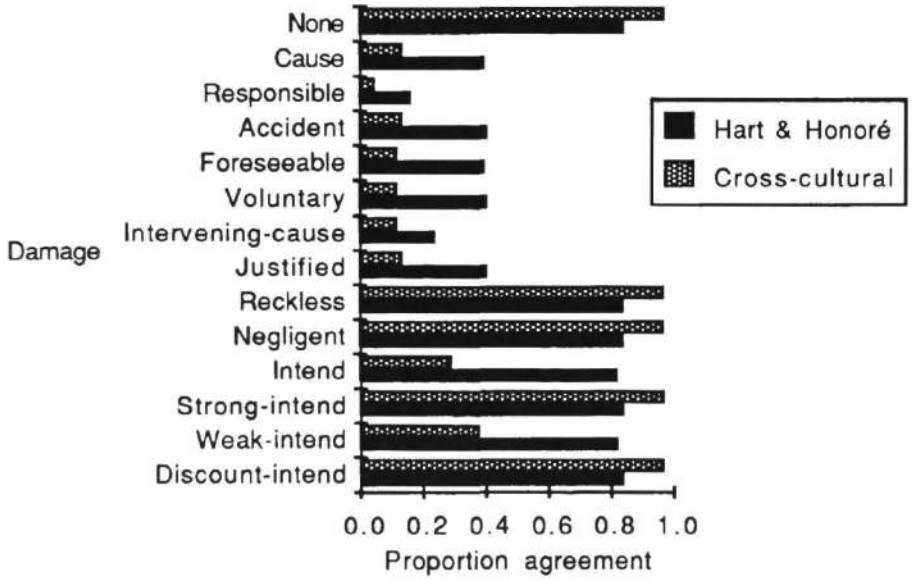

Figure 2. Proportion agreement with judicial decisions after rule damage.

The results for the 58 cross-cultural cases are likewise presented in hatched bars in Figure 2 in terms of proportion agreement with the actual decision. The proportion agreement produced by each type of damage was again contrasted with that produced by the fully intact model using log-linear analysis. The results were similar to those for the Hart and Honoré cases with two striking exceptions -- *intend* and *weak-intend*. The *intend* rule concludes that the harm was intended only if there is either strong or weak evidence of intention. The *weak-intend* rule concludes that there is weak evidence of intention if the accused's actions were neither reckless nor negligent, and one or more of the following is true: the accused's intending the harm cannot be discounted, the accused monitored his actions, or the harm benefitted the accused. A *strong-intend* rule concludes that there is strong evidence for intention if either the harm was described as intentional or the following all hold: the accused's plan was known, the accused's plan included the harm, and the harm was caused as planned.

Damage to the *intend* rule produced a proportion agreement of .82 in the Hart and Honoré cases, but only .29 for the cross-cultural cases, $X^2 = 37.32$, $p < .001$. Analogously, damage to the *weak-intend* rule produced a proportion agreement of .82 in the Hart and Honoré cases, but only .38 for the cross-cultural cases, $X^2 = 28.05$, $p < .001$. For the cross-cultural cases, damage to each rule did significantly lower the agreement rate as compared to the intact model (mean = .17, all $ps < .001$), except for the *reckless*, *negligent*, *strong-intend*, and *discount-intend* rules (mean = .97, all $ps > .5$).

This initially surprising discrepancy between the western and traditional cases can be explained by recalling that the latter cases are conceptually much simpler than the former. The Hart and Honoré cases were typically subtle and complex, turning on issues such as causation, intervening causation, negligence, or recklessness. In contrast, the cross-cultural cases were extremely straightforward, typically involving weak evidence for intention of the accused and not hinging on difficult issues such as causation, intervening causation, or alternative mental states such as negligence or recklessness. A typical cross-cultural case involved, for example, one person stealing another person's pig, and either making or not making restitution for it. Because there was no alternative mental state to compensate for damage to the *weak-intend* and *intend* rules in these latter cases, this type of rule-damage was fatal to proportion of agreement. In contrast, damage to the other mental state rules had little effect on proportion agreement since these other rules were rarely relevant to the cases.

CONCLUSIONS AND DISCUSSION

The simulations show that the MR program is computationally sufficient to qualitatively match human reasoning about harm-doing. They further suggest that the MR model and the psychological theory on which it is based have a large degree of cultural universality. Cultures undoubtedly vary in their value judgments about what constitutes what degree of harm and what sorts of justifications outweigh harms, but they do not appear to differ in the rules they apply to moral judgments about harm-doing. This universality could have important implications for explaining the development of these moral judgment rules.

The construction of the MR program was extremely useful in forcing a rigorous specification of an increasingly complex psychological theory. This was particularly true in terms of specifying how different parts of the theory ought to interact. Early simulations identified a number of inconsistencies and anomalies in the rule base. Corrections of these problems were then tested in subsequent simulations. It would have been extremely difficult to identify and test these issues without the benefit of a working computational model.

One unanticipated sort of interaction among of parts of the model was the degree to which mental state rules compensate for each other. Moreover, the extent of this compensation was found to interact with the subtlety of the case. For the relatively difficult cases found in books on western jurisprudence, the mutual compensation of mental state rules was extensive since many of the mental state rules were relevant to many of these cases. But for the relatively simple cases found in anthropological reports from traditional cultures, this mutual compensation disappeared since only a few of the mental state rules were relevant.

LIMITATIONS OF THE CURRENT MODEL AND SUGGESTIONS FOR FUTURE WORK

Although encouraging, these simulations do not constitute a very complete test of MR. They compare only one decision, blameworthiness, a decision that can be reached by a number of different paths through the rule base. More refined and more complete tests of

MR could be made by examining the intermediate decisions of ordinary subjects reading vignettes of legal cases.

Much more work is needed on the process of encoding the initial description of the case. Currently, the programmer translates the English version of the case into an attribute-value frame that the rules can use to make further inferences. It is likely, however, that subjects differ substantially in how they encode and interpret at least some cases. Such encoding differences would undoubtedly lead to differing conclusions about the case.

RELATED WORK

Pennington and Hastie (1988) investigated decision processes in simulations of trials by jury and constructed a theory of how jurors organize the information emerging during a trial. It is their view that the juror's main task is to construct a causal explanation of how the harm was produced and that the judge, through final instructions to the jury, often provides the responsibility and blame rules necessary to reach a decision on guilt or innocence. The emphasis of the present project is on explicating these rules as used implicitly by ordinary reasoners.

Thagard (1989) applied a connectionist model of explanatory coherence, ECHO, to two legal cases in which the prosecution and the defense advocated incompatible ways of explaining the evidence. Given a network of coherence and incoherence relations among propositions describing the evidence and competing claims of the case, ECHO propagated activation across the network to maximize the coherence of the network. ECHO focuses only on the issue of how the harm was caused.

Bain's JUDGE (in Riesbeck & Schank, 1989) is a case based reasoning program in the area of criminal sentencing. It uses rules to build up its case library. It has been our experience (via think aloud protocols and pointed questions) that, perhaps unlike judges, ordinary people have no extensive case libraries to draw on.

REFERENCES

Darley, J. M., & Shultz, T. R. (1990). Moral rules: Their content and acquisition. *Annual Reviews of Psychology*, **41**, 525-556.

Hart, H. L. A., & Honoré, A. M. (1959). *Causation in the law*. Oxford: Clarendon Press.

Pennington, N., & Hastie, R. (1988). Explanation-based decision making: Effects of memory structure on judgment. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **14**, 521-533.

Pospisil, L. (1958). *Kapauku Papuans and their law*. New Haven: Yale University Press.

Riesbeck, C. K., & Schank, R. C. (1989). *Inside case-based reasoning*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Shultz, T. R., & Schleifer, M. (1983). Towards a refinement of attribution concepts. In J. Jaspars, F.D. Fincham, & M. Hewstone (Eds.), *Attribution theory and research: Conceptual, developmental and social dimensions* (pp. 37-62). London: Academic Press.

Thagard, P. (1989). Explanatory coherence. *Behavioral and Brain Sciences*, **12**, 435-467.