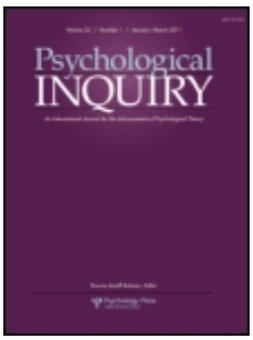

Routledge
Taylor & Francis Group

Psychological Inquiry

An International Journal for the Advancement of Psychological Theory

ISSN: 1047-840X (Print) 1532-7965 (Online) Journal homepage: https://www.tandfonline.com/loi/hpli20

The Prototype Model of Blame: Freeing Moral Cognition From Linearity and Little Boxes

Chelsea Schein & Kurt Gray

To cite this article: Chelsea Schein & Kurt Gray (2014) The Prototype Model of Blame: Freeing Moral Cognition From Linearity and Little Boxes, Psychological Inquiry, 25:2, 236-240, DOI: 10.1080/1047840X.2014.901903

To link to this article: https://doi.org/10.1080/1047840X.2014.901903

Psychological Inquiry, 25: 236–240, 2014

ISSN: 1047-840X print / 1532-7965 online
DOI: 10.1080/1047840X.2014.901903

Routledge
Taylor & Francis Group

The Prototype Model of Blame: Freeing Moral Cognition From Linearity and Little Boxes

Chelsea Schein and Kurt Gray
Department of Psychology, University of North Carolina, Chapel Hill, Chapel Hill, North Carolina

Little boxes on the hillside, little boxes made of ticky tacky, little boxes on the hillside, little boxes all the same. — Malivna Reynolds

The song "Little Boxes" describes a cookie-cutter town in which the path to success is both precise and invariant: School leads to summer camp, which leads to university, then to prestigious careers, then to spouses and children, and finally to country club memberships. Whether this song describes perfection or perdition is debatable, but we can all agree that the path of life is never so straight. People get pregnant at summer camp, get married in university, and get laid off from their prestigious jobs. Despite the appeal of discrete life stages, real life is messy, with blurred boundaries, back-tracking, dead-ends, and many loops. Like this song, Malle, Guglielmo, and Monroe (this issue) describe a model in which blame is divided into discrete little boxes, linked by an invariant sequence. Just as with real life, we suggest that moral cognition cannot be confined to discrete boxes, whether in structure or in sequence. Instead of a static linear path, recent research suggests that blame judgment proceeds like a swirling vortex, pulling together cognitive elements toward an underlying prototype (e.g., Gray & Schein, 2012). This messier and more dynamic view of judgment is suggested by modern multilevel understandings of the mind and by two phenomena not discussed by Malle et al. (this issue)—dyadic completion and moral typecasting. We explore these arguments and conclude that judgments of blame are best explained by fuzzy prototypes, not by strict paths.

Prototypes not Paths

Malle et al. (this issue) advance a model of blame in which the mind acts as a simple computer, calculating blame through an invariant sequence of binary logic gates: Agent Causality? Yes. Intention? Yes. Mitigating reasons? No. Then, output = Blame. Similar "if–then" models of the mind, with discrete paths and logic branches, were advanced decades ago by Strong AI (Turing, 1950), and although much was learned from this research, its most important lesson was that the mind defied such simple modeling (Dreyfus, 1979, 1981, 2007). Cognitive processes from visual perception (Anderson, Silverstein, Ritz, & Jones, 1977; Spivey & Dale, 2006) to social categorization (Corneille, Hugenberg, & Potter, 2007) are too dynamic and complex to be understood by rigid path models. Instead, cognition is better modeled by multilevel neural nets that allowed extensive feedback loops and powerful top-down constraints (Rumelhart & McClelland, 1986).

In categorization judgments, top-down constraints are afforded by prototypes, which are cognitive structures that represent key features and/or canonical cases of categories (Rosch, 1978). For example, a prototype of "birds" would be something that looks like a robin or sparrow, and categorization judgments depend on the overall similarity between this prototype and specific examples. Thus, a chickadee would be more robustly categorized than a penguin. Although both prototype and path models use similar criteria for categorization (for birds: feathers, beaks, flight), prototypes judgments are holistic, whereas path judgments use strict binary if–then rules. The strict sequential rules of path models draw firmer category boundaries than prototype models—which allow for degrees of matching—and these firm boundaries inevitably lead to miscategorization. For example, a path model based on feathers, beaks, and flight would fail to categorize penguins as birds. One might think that path models just need right rules to be accurate, but philosophers and psychologists have long recognized the impossibility of specifying the necessary and sufficient features of any complex category, whether birds or blame (Wittgenstein, 2001). In every domain of research in which they are compared, prototype models predict human judgment better than path models (Spivey & Dale, 2004; Thelen, 1996), and we suggest that the same is true with blame.

If blame judgments involve a prototype, what are its features? Malle et al. (this issue) have identified some features—cause, an agent, and intentionality—that have been suggested to form a universal moral grammar, in which "INTENT + CAUSE + HARM = WRONG" (Mikhail, 2007). Synthesizing this insight with classic work in cognitive psychology, a new theory of morality suggests that the prototype of wrongdoing is dyadic, consisting of two perceived minds: an intentional agent harming a suffering

patient (Gray, Waytz, & Young, 2012; Gray, Young, & Waytz, 2012). This moral dyad of agent + patient grows from the evolutionary significance of harm, the dyadic structure of causation and language (Brown & Fish, 1983; Strickland, Fisher, & Knobe, 2012), and the deep roots of empathy that lends affective power to perceived suffering (Davis, 1996; Preston & de Waal, 2001). Consistent with a dyadic prototype of morality, acts that closely resemble this dyad (e.g., murder) are most robustly categorized as instances of immorality (Schein & Gray, 2013).

As the moral dyad predicts the immorality of *acts*, some may suggest that it is tangential to blame, which concerns judgments of *agents*. Although it is true that the blameworthiness of agents can be conceptually distinguished from the wrongness of acts (Cushman, 2008), in practice these judgments are very highly correlated. More important, the dyadic prototype contains a *moral agent*, to whom blame can be assigned. We suggest that judgments of blame reflect not a qualitatively different mode of judgment but simply an agent-centric understanding of the dyad, in which participants weigh agent-relevant features (e.g., intention) more heavily than patient-relevant features (e.g., suffering). In both wrongness and blame judgments, the underlying prototype is constant but blame involves a slight shift in focus.

We suggest that the distinction between blame and wrongness further disappears in intuitive judgments. Although someone can theoretically believe that gay sex is immoral without blaming the lovers, these two judgments are almost always conflated and require cognitive resources and explicit reasoning to separate (Haidt & Hersh, 2001). Predicating a model of blame on a firm boundary between blame/wrongness fails to acknowledge the overwhelming overlap between these two concepts, especially in typically quick and intuitive moral cognition (Haidt, 2001). In contrast to path models, a dyadic prototype predicts fuzzy lines between wrongness and blame, and among intention, causation and suffering. Rather than the separate little boxes of a path model, a prototype model predicts overlap and even mutual activation of intention, causation and suffering. See Figure 1. Where there are some features of blame, people assume the presence of others.

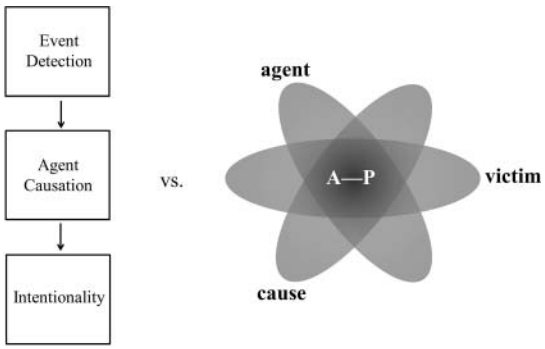

Figure 1. The path model (left) vs. the blame prototype (right) of blame. *Note.* In the prototype model, the moral dyad (agent harming patient) is the gravitational center of blame.

Figure 2. This image consists of black splotches, but due to top-down conceptual knowledge influencing perception, you likely just see a panda.

Cognitive Coherence and Dyadic Completion

The moral dyad is not merely a static representation of immorality; prototypes exert powerful top-down influences on cognition, bending perceptions to align with itself. This phenomenon is called coherence because prototypes make various elements form a coherent whole (Smith, 1996). For example, when the prototype/stereotype of "Black men" is activated, it leads people to see targets as athletic, aggressive, lazy, *and* musical, even if none of those traits are applicable (Lepore & Brown, 1997; Smith & Zarate, 1990). Coherence also exists in low-level visual phenomenon, such as in Figure 2 when the prototype of "panda" leads you to see a coherent animal rather than disparate splotches of ink (Humphrey, 1924). Another way to think about coherence is as "cognitive gravity,"[1] as perceptions are pulled and distorted by the central prototype. A prototype model of blame suggests that the elements of intention and causation are not merely objectively evaluated inputs—as the path model predicts—but are themselves shaped by the influence of the dyadic prototype.

[1]More technically, the prototype acts as an attractor basin (Spivey & Dale, 2006).

Although not discussed by Malle et al. (this issue), multiple studies reveal the top-down influence of the moral dyad (Gantman & Van Bavel, in press; Gray, Schein, & Ward, in press). This process of *dyadic completion* means that *all* elements of the dyad are activated—intentional agent, causation, suffering patient—when *any* of them are activated with even the slightest whiff of immorality. If you think praying for the death of your mother is wrong (immoral agent), then you infer the presence of causation and harm (e.g., she will be in an accident and die), and therefore blame. If you think homosexuality is wrong, then you perceive it to destroy society (Bryant, 1977) and to make children suffer (Gray et al., in press). In other words, full blown judgment of blames can occur without "objectively" meeting the initial criteria in the path model, because the prototype shapes perceptions of those initial criteria.

Most relevant to blame is *agentic dyadic completion* (Gray et al., in press), in which the simple perception of undeserved suffering can compel the perception of a blameworthy agent, despite their ambiguous involvement. The devastation of Hurricane Katrina left many people pointing fingers at FEMA or President Bush or even the gay-pride parade that was scheduled to occur days after Hurricane Katrina (Gross, 2008). Strikingly, a recent poll even found that some Americans currently blame President Obama for the poor governmental response to the disaster (who took office *after* the hurricane; Public Policy Polling, 2013). When people see unjust suffering, they cannot help but see a blameworthy agent to complete the moral dyad—whether in other people, animals, or God (Gray & Wegner, 2010a). The ability for judgments of immorality and blame to feedback and influence the perceptions of so-called inputs cannot be easily explained by the path model. Certain mental state inferences lead to judgments of wrongness, but wrongness judgments also lead to mental state inferences (Knobe, 2003). Rather than a linear path, we suggest that the geometry of blame is a spiral in which one element is activated (e.g., immoral act, or undeserved suffering), and activates all other elements as it swirls in toward the central dyadic prototype. See Figure 3.

The coherence process of dyadic completion can be motivational: Blaming God for the tragic death of a young child can furnish a sense of meaning, and seeing intentional wrongs as more harmful (Ames & Fiske, 2013) or more painful (Gray, 2012; Gray & Wegner, 2008) can better allow us to punish transgressors. However, coherence processes can also operate automatically without obvious motivation. In Figure 2, you see a panda without first feeling a deep burning urge for panda perception, and in dyadic completion, people see agents and patients simply by virtue of cognitive coherence. In their article,

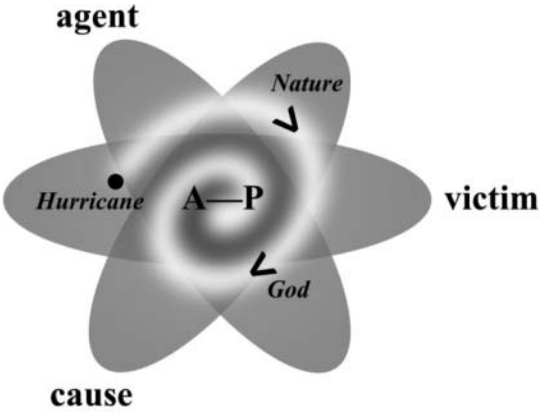

Figure 3. An example of agentic dyadic completion. *Note.* The suffering caused by Hurricane Katrina leads people to see an agent to blame for that suffering. More broadly, activating *any* element of the dyad, activates *all* elements because of the gravitational pull of the moral dyad.

Malle et al. (this issue) discount some demonstrations of top-down influences in blame judgments (e.g., Alicke, 2000; Knobe, 2003) because these top-down effects may involve motivation. The prototype model of blame completely avoids this criticism because the moral dyad can shape perceptions without motivation; dyadic completion occurs in the span of milliseconds and does not require the activation of any specific goals (Gray et al., in press). Thus, dyadic completion is one phenomenon inconsistent with a path model, but is consistent with a prototype model—moral typecasting is another one.

Moral Typecasting: Victims Escape Blame

Are you less blameworthy for murder if bullied as a child? The path model suggests not. Unless past victimization can directly justify an immoral act (e.g., killing an abusive spouse), the path model provides no route by which previous victimhood can reduce blame. It is clear, however, that even completely irrelevant previous victimization *does* decrease judgments of blame (Gray & Wegner, 2009, 2010b, 2011b). Someone who suffered years ago is ascribed less blame than someone who did not suffer, even if they both completed the exact same action (Gray & Wegner, 2011b). Although the path model cannot account for the mitigating effect of irrelevant victimhood, a prototype model can.

Dyadic morality suggests the template of wrongness/blame is two different people as agent and patient (Gray, Waytz, et al., 2012). Because people are typically either an agent or a patient in specific moral acts, we *generally* see others as *either* moral agents (heroes/villains) *or* moral patients (victims/beneficiaries)—a phenomenon called moral typecasting (Gray & Wegner, 2009). Moral typecasting predicts that the more you are seen as a victim, the less you are seen as

a villain—not because of any motivational processes but because victims just don't seem like villains. Just as it is hard to imagine a robot as alive (Gray & Wegner, 2012), or a porn star as a CEO (Gray, Knobe, Sheskin, Bloom, & Barrett, 2011), studies reveal that it is hard to imagine a suffering victim perpetrating evil (Gray & Wegner, 2011b).

Blame and Deliberative Adjustments of Blame

Unlike a prototype model of blame, a path model cannot account for moral typecasting and dyadic completion. However, one could argue that this is inconsequential, because although moral dyad operates at the level of intuition (Gray et al., in press) and emotion (Gray & Wegner, 2011a), the path model concerns conscious deliberation and explicit justifications. In the courtroom, judges and juries must logically calculate whether blame is warranted, and these judgments are clearly important—in fact, they can be a matter of life or death. The path model nicely outlines how such explicit reasoning could progress, but ample research suggests that everyday judgments seldom involve explicit reasoning and instead arise from intuition (Greenwald & Banaji, 1995; Haidt, 2001). In particular, most moral judgments are intuitive and emotional (Haidt, 2001), and so the everyday applicability of the path model is unclear. Indeed, the only unambiguous evidence for the path model comes from experimenter-led verbal protocols that encourage conscious reason and deliberate rationalization. Moreover, explicit reasoning likely operates upon intuitive moral judgments rather than creating them de novo, suggesting that the path model is best understood as an adjustment protocol upon judgments generated via a dyadic prototype: first the dyad, and then path-based adjustments and justifications.

The path model is an elegant theory that clearly outlines how people *should* assign blame in legal settings—and perhaps how people can assign blame given maximum motivation and resources—but we suggest that the prototype model best predicts how people do actually assign blame. Dyadic completion shows that people are only too happy to blame others without warrant, punishing scapegoats including children (Hill, 2002), pigs (Oldridge, 2004), and God (Gray & Wegner, 2010a). Moral typecasting also shows that people are happy to excuse others from blame without warrant, allowing the past suffering of transgressor to shape their judgments (Gray & Wegner, 2009). These two phenomena show that blame, like life, is perhaps too messy to be understood by straight lines between little boxes. Although clearly marked paths represent an important ideal, experience tells us the routes to anything—whether truth, salvation, or happiness—are full of loops and switchbacks. The most direct route to understanding blame may not be a straight path but instead a spiral that bends ever towards the dyad.

Note

Address correspondence to Kurt Gray, Department of Psychology, University of North Carolina, Chapel Hill, Chapel Hill, NC 27599. E-mail: kurtgray@unc.edu

References

Alicke, M. D. (2000). Culpable control and the psychology of blame. *Psychological Bulletin*, *126*, 556–574.

Ames, D. L., & Fiske, S. T. (2013). Intentional harms are worse, even when they're not. *Psychological Science*, *24*, 1755–1762.

Anderson, J. A., Silverstein, J. W., Ritz, S. A., & Jones, R. S. (1977). Distinctive features, categorical perception, and probability learning: Some applications of a neural model. *Psychological Review*, *84*, 413.

Brown, R., & Fish, D. (1983). The psychological causality implicit in language. *Cognition*, *14*, 237–273. doi:10.1016/0010-0277(83)90006-9

Bryant, A. (1977). *The Anita Bryant story: The survival of our nation's families and the threat of militant homosexuality*. Grand Rapids, MI: Revell.

Corneille, O., Hugenberg, K., & Potter, T. (2007). Applying the attractor field model to social cognition: Perceptual discrimination is facilitated, but memory is impaired for faces displaying evaluatively congruent expressions. *Journal of Personality and Social Psychology*, *93*, 335–352.

Cushman, F. (2008). Crime and punishment: Distinguishing the roles of causal and intentional analyses in moral judgment. *Cognition*, *108*, 353–380. doi:10.1016/j.cognition.2008.03.006

Davis, M. H. (1996). *Empathy: A social psychological approach*. Boulder, CO: Westview Press.

Dreyfus, H. L. (1979). *What computers can't do: The limits of artificial intelligence* (Vol. 1972). New York, NY: Harper & Row. Retrieved from http://www.getcited.org/pub/101997805

Dreyfus, H. L. (1981). From micro-worlds to knowledge representation: AI at an impasse. In J. Haugeland (Ed.), *Mind design II* (pp. 161–204). Cambridge, MA: MIT Press.

Dreyfus, H. L. (2007). Why Heideggerian AI failed and how fixing it would require making it more Heideggerian. *Philosophical Psychology*, *20*, 247–268.

Gantman, A. P., & Van Bavel, J. J. (in press). The moral pop-out effect: Enhanced perceptual awareness of morally relevant stimuli. *Cognition*. Retrieved from http://www.psych.nyu.edu/vanbavel/lab/documents/Gantman.VanBavel.MoralPopoutCognition2014.pdf

Gray, K. (2012). The power of good intentions: Perceived benevolence soothes pain, increases pleasure, and improves taste. *Social Psychological and Personality Science*, *3*, 639–645.

Gray, K., Knobe, J., Sheskin, M., Bloom, P., & Barrett, L. F. (2011). More than a body: Mind perception and objectification. *Journal of Personality and Social Psychology*, *101*, 1207–1220.

Gray, K., & Schein, C. (2012). Two minds vs. two philosophies: Mind perception defines morality and dissolves the debate between deontology and utilitarianism. *Review of Philosophy and Psychology*, *3*, 405–423. doi:10.1007/s13164-012-0112-5

Gray, K., Schein, C., & Ward, A. F. (in press). The myth of harmless wrongs in moral cognition: Automatic dyadic completion from sin to suffering. *Journal of Experimental Psychology: General*.

Gray, K., Waytz, A., & Young, L. (2012). The moral dyad: A fundamental template unifying moral judgment. *Psychological Inquiry*, *23*, 206–215.

Gray, K., & Wegner, D. M. (2008). The sting of intentional pain. *Psychological Science*, *19*, 1260–1262. doi:10.1111/j.1467-9280.2008.02208.x

Gray, K., & Wegner, D. M. (2009). Moral typecasting: Divergent perceptions of moral agents and moral patients. *Journal of Personality and Social Psychology*, *96*, 505–520. doi:10.1037/a0013748

Gray, K., & Wegner, D. M. (2010a). Blaming God for our pain: Human suffering and the divine mind. *Personality and Social Psychology Review*, *14*, 7–16. doi:10.1177/1088868309350299

Gray, K., & Wegner, D. M. (2010b). Torture and judgments of guilt. *Journal of Experimental Social Psychology*, *46*, 233–235. doi:10.1016/j.jesp.2009.10.003

Gray, K., & Wegner, D. M. (2011a). Dimensions of moral emotions. *Emotion Review*, *3*, 227–229.

Gray, K., & Wegner, D. M. (2011b). To escape blame, don't be a hero—Be a victim. *Journal of Experimental Social Psychology*, *47*, 516–519. doi:10.1016/j.jesp.2010.12.012

Gray, K., & Wegner, D. M. (2012). Feeling robots and human zombies: Mind perception and the uncanny valley. *Cognition*, *125*, 125–130. doi:10.1016/j.cognition.2012.06.007

Gray, K., Young, L., & Waytz, A. (2012). Mind perception is the essence of morality. *Psychological Inquiry*, *23*, 101–124.

Greenwald, A. G., & Banaji, M. R. (1995). Implicit social cognition: Attitudes, self-esteem, and stereotypes. *Psychological Review*, *102*, 4.

Gross, K. (2008, May 16). Paster John Hagee on Christian Zionism, Katrina. National Public Radio. Retrieved from http://www.npr.org

Haidt, J. (2001). The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychological Review*, *108*, 814–834.

Haidt, J., & Hersh, M. A. (2001). Sexual morality: The cultures and emotions of conservatives and liberals. *Journal of Applied Social Psychology*, *31*, 191–221. doi:10.1111/j.1559-1816.2001.tb02489.x

Hill, F. (2002). *A delusion of Satan: The full story of the Salem witch trials*. New York, NY: Da Capo Press.

Humphrey, G. (1924). The psychology of the gestalt. *Journal of Educational Psychology*, *15*, 401–412. doi:10.1037/h0070207

Knobe, J. (2003). Intentional action and side effects in ordinary language. *Analysis*, *63*, 190–193.

Lepore, L., & Brown, R. (1997). Category and stereotype activation: Is prejudice inevitable? *Journal of Personality and Social Psychology*, *72*, 275–287. doi:10.1037/0022-3514.72.2.275

Mikhail, J. (2007). Universal moral grammar: Theory, evidence and the future. *Trends in Cognitive Sciences*, *11*, 143–152. doi:10.1016/j.tics.2006.12.007

Oldridge, D. J. (2004). *Strange histories: The trial of the pig, the walking dead, and other matters of fact from the medieval and renaissance worlds*. New York, NY: Routledge.

Preston, S. D., & de Waal, F. B. M. (2001). Empathy: Its ultimate and proximate bases. *Behavioral and Brain Sciences*, *25*, 1–20. doi:10.1017/S0140525X02000018

Public Policy Polling. (2013, August 21). *In Louisiana, Clinton keeps up, governor falls—Public Policy Polling*. Retrieved from http://www.publicpolicypolling.com/main/2013/08/in-louisiana-clinton-keeps-up-governor-falls.html

Rosch, E. (1978). Principles of categorization. In E. Rosch & B. B. Lloyd (Eds.), *Cognition and categorization* (pp. 251–270). Hillsdale, NJ: Erlbaum.

Rumelhart, D. E., & McClelland, J. L. (1986). *Parallel distributed processing: Explorations in the microstructure of cognition, Volumes 1 and 2*. Cambridge, MA: MIT Press.

Schein, C., & Gray, K. (2013). *Thou shalt not harm: Victims are central across moral diversity*. Manuscript under review.

Smith, E. R. (1996). What do connectionism and social psychology offer each other? *Journal of Personality and Social Psychology*, *70*, 893–912. doi:10.1037/0022-3514.70.5.893

Smith, E. R., & Zarate, M. A. (1990). Exemplar and prototype use in social categorization. *Social Cognition*, *8*, 243–262. doi:10.1521/soco.1990.8.3.243

Spivey, M. J., & Dale, R. (2004). On the continuity of mind: Toward a dynamical account of cognition. In *Psychology of learning and motivation* (Vol. 45, pp. 87–142). New York, NY: Academic Press. Retrieved from http://www.sciencedirect.com/science/article/pii/S0079742103450032

Spivey, M. J., & Dale, R. (2006). Continuous dynamics in real-time cognition. *Current Directions in Psychological Science*, *15*, 207–211.

Strickland, B., Fisher, M., & Knobe, J. (2012). Moral structure falls out of general event structure. *Psychological Inquiry*, *23*, 198–205. doi:10.1080/1047840X.2012.668272

Thelen, E. (1996). *A dynamic systems approach to the development of cognition and action*. Cambridge, MA: MIT Press. Retrieved from http://books.google.com/

Turing, A. M. (1950). Computing machinery and intelligence. *Mind*, *59*, 433–460.

Wittgenstein, L. (2001). *Philosophical investigations: The German text, with a revised English translation* (G. E. M. Anscombe, Trans.). Malden, MA: Blackwell.