

RATIONALITY AND SOCIAL RESPONSIBILITY

Essays in Honor of Robyn Mason Dawes

Edited by
Joachim I. Krueger



 Psychology Press
Taylor & Francis Group

Psychology Press
Taylor & Francis Group
270 Madison Avenue
New York, NY 10016

Psychology Press
Taylor & Francis Group
27 Church Road
Hove, East Sussex BN3 2FA

© 2008 by Taylor & Francis Group, LLC

Printed in the United States of America on acid-free paper
10 9 8 7 6 5 4 3 2 1

International Standard Book Number-13: 978-0-8058-5996-6 (Hardcover)

Except as permitted under U.S. Copyright Law, no part of this book may be reprinted, reproduced, transmitted, or utilized in any form by any electronic, mechanical, or other means, now known or hereafter invented, including photocopying, microfilming, and recording, or in any information storage or retrieval system, without written permission from the publishers.

Trademark Notice: Product or corporate names may be trademarks or registered trademarks, and are used only for identification and explanation without intent to infringe.

Library of Congress Cataloging-in-Publication Data

Rationality and social responsibility : essays in honor of Robyn Mason Dawes / editor,
Joachim I. Krueger.

p. cm. -- (Modern pioneers in psychological science) (An APS-LEA series)
Includes bibliographical references.

ISBN 978-0-8058-5996-6 (alk. paper)

1. Reasoning (Psychology) 2. Thought and thinking. 3. Decision making. 4.
Responsibility. I. Dawes, Robyn M., 1936-. II. Krueger, Joachim I.

BF442.R37 2008
153.4--dc22

2007045940

Visit the Taylor & Francis Web site at
<http://www.taylorandfrancis.com>
and the Psychology Press Web site at
<http://www.psypress.com>

I

A PSYCHOLOGIST BETWEEN LOGOS AND ETHOS

Joachim I. Krueger
Brown University

At the 17th Annual Convention of the Association for Psychological Science, Los Angeles, May 29, 2005, Robyn Mason Dawes was honored with a Festschrift conference. Following a convivial banquet on the eve of the Festschrift symposium, several of Robyn's friends and colleagues presented papers on recent research that was, in one way or another, indebted to his intellectual inspiration over the years. This volume is composed of chapter-length reports of most of these research enterprises as well as some chapters authored by individuals who did not speak at the symposium. With characteristic modesty, Robyn requested that the contributors present their finest work and refrain from emphasizing his contributions to these efforts. It falls on this introductory chapter to provide the overarching context for these lines of research and a sense of Robyn's abiding influence.

It is fitting to begin with a biographical note. Robyn entered the graduate program in clinical psychology at the University of Michigan in 1958. His stint in that program included some experiences that shaped his outlook on psychological science for years to come. Robyn noticed that clinical psychology was caught between the two frames of being seen as an art and a science. In his view, the latter was neglected. He found that scientific advances showing the limits of clinical diagnosis were being ignored. In one poignant instance, Robyn was asked to administer a Rorschach test to a 16-year old girl who had been admitted because of sexual relations with an older man, which had led to

strained relations with her family. After analyzing the Rorschach protocol, Robyn found that just one of her responses to any of the Rorschach cards was "abnormal" (i.e., "poor form"). At a case conference, he argued that this low rate of nonstandard responding did not warrant a diagnosis of schizophrenia. In essence, his argument was that her performance was no different than what would be expected from base rate responding in the general population. If anything, her responses were more "reality oriented." Robyn was overruled by experienced clinicians, who rejected his statistical argument. Mr. Dawes, he was told, you may understand mathematics, but you do not understand psychotic people ("Just look at the card. Does it look like a bear to you? She must have been hallucinating."). With that, Robyn left the program and turned his attention to mathematical psychology.

In due course, his collaboration with his mentor Clyde Coombs and his cohort Amos Tversky led to the publication of the classic monograph of this field (Coombs, Dawes, & Tversky, 1970). Much of his own empirical work was henceforth dedicated to the detection of psychological distortions in human judgment (Dawes, 2001). From the outset, he did not single out clinical experts as victims of irrational thought but tried to understand general principles of mind that yield systematic biases in judgment. With this orientation, he placed himself in the tradition of Jerome Bruner and his colleagues, who had recently laid the groundwork for the cognitive revolution in psychology. Reaching further back still, the influence of Bartlett's (1932) work on the role of narratives in remembering turned out to be relevant.

To Robyn, human reasoning occurs in both a narrative and a scientific mode. Narratives provide meaning, help remembering, and simplify otherwise complex quantitative input, yet these same narratives can systematically interfere with logic or statistical rationale. In his dissertation, Robyn presented his participants with declarative sentences representing a suite of set relations (Dawes, 1964, 1966). Some relations were nested, such as when all X were Y (inclusion) or when no X were Y (exclusion). Other relations were disjunctive, such as when some X were Y. Immediately after presentation, Robyn probed the participants' recollections of these relations. As he had suspected, most participants were more likely to remember disjunctive relations as being nested than to remember nested relations as being disjunctive (and were more confident in their recall when it involved this error than when it was correct). In other words, he had not only identified systematic memory distortions, but he had correctly predicted which type of distortion was more likely to occur. From a narrative point of view, these distortions were benign in that they yielded mental representations that were sim-

pler than the reality they meant to portray. From a scientific or paradigmatic point of view, however, these distortions were worrisome because they could lead to incoherent judgments and ultimately harmful action. As Robyn cautioned, "we may think of these people as one-bit minds living in a two-bit world" (Dawes, 1964, p. 457).

Misremembering disjunctive relations as being nested is an error of overgeneralization, a kind of error Robyn had already seen in action. Giving an "abnormal" response to a particular Rorschach card and being psychotic are instances of disjunctive sets. To say that everyone who gives this response is psychotic is to claim that the former set is nested within the latter. In time, Robyn approached these types of judgment task from a Bayesian point of view. Harkening back to his earlier experience with clinical case conferences, it could be said that the clinicians began with the idea that there was a high probability of finding evidence for an offending response if the person was ill (i.e., $p[E|I]$). Their diagnostic judgment, however, was the inverse of this conditional probability, namely, the probability that the person was ill given that one abnormal response was in evidence (i.e., $p[I|E]$). Meehl and Rosen (1955) had shown that base rate neglect pervades human judgment in general and clinical judgment in particular. Inasmuch as the base rate of making the response, $p(E)$, is higher than the base rate of being ill, $p(I)$, the probability of illness given the evidence from testing can be very low indeed.

Along with his friends Amos Tversky and Daniel Kahneman, Robyn stimulated the growth of the psychology of judgment and decision making. Whereas Tversky and Kahneman documented the base rate fallacy in a series of widely noted experiments (Tversky & Kahneman, 1974), Robyn discussed numerous examples of non-Bayesian thinking in his writing. His 1988 book, *Rational Thought in an Uncertain World* (see also Hastie & Dawes, 2001, for a revised 2nd edition; a 3rd edition is in preparation), is a classic exposition of common impediments to rationality. His critique of clinical psychology and psychotherapy (*House of Cards* [Dawes, 1994]) is the authoritative application of the principles of judgment and decision making to that field. Robyn's empirical work yielded a report in which he and his collaborators presented a rigorous correlational test of the overgeneralization hypothesis (Dawes, Mirels, Gold, & Donahue, 1993). Unlike Tversky and Kahneman, who provided their participants with base rate information, Robyn and his colleagues asked their participants to estimate all four constituent Bayesian probabilities themselves. This within-person method yielded strong evidence for overgeneralization. When estimating $p(I|E)$, participants neglected the base rates of $p(E)$ and $p(I)$ that they themselves had estimated.

Instead, they appeared to derive $p(I|E)$ directly from $p(E|I)$. Bayes's theorem states that the two inverse conditional probabilities are the same only when the two base rates are the same (which they were not for most participants). In Robyn's analysis, people assumed a symmetry of association that is rare in the real world.

Associations between variables are often represented as correlations. One may want to predict, for example, illness versus health from positive versus negative test results, yet a perfect correlation can only be obtained when the two base rates are the same. Overgeneralization becomes more likely inasmuch as the base rate of the diagnostic sign is larger than the base rate of the underlying illness. It is worth considering a numerical example. Suppose a certain test response is rather common ($p[E] = .8$), whereas the underlying illness is rare ($p[I] = .3$). The ceiling for the correlation between the two is .327. Even if everyone who is ill emits the response (i.e., $p[E|I] = 1$), the inverse probability, $p(I|E)$, is only .375. The former probability expresses the *sensitivity* of the test, or how typical the response is of the underlying disposition. The inverse probability, however, also depends on the ratio of the base rates. According to Bayes's theorem, $p(I|E) = p(E|I)p(I)/p(E)$, where $p(E) = p(I)p(E|I) + p(-I)p(E|-I)$. In other words, $p(E)$ depends greatly on the probability of the evidence given the *absence* of the illness (which, in turn, is the complement of the test's *specificity*, or $1 - p[-E|-I]$). This is what people neglect to take into account and what leads them to confuse pseudo-diagnosticity with true diagnosticity.

As the predictive validity of a test can be expressed as the correlation between test results and actual health status, this Bayesian analysis shows that the ceiling of validity coefficients becomes lower as the two base rates become more discrepant. In Figure 1.1 the maximum correlations are plotted against the ratio of the base rates. Data for this illustration were constructed by letting $p(I)$ range from .05 to .95 and letting $p(E)$ range from .05 to $p(I)$. Across the simulated instances, the correlation between the base rate ratio and their maximum statistical association is $-.65$. When the natural logarithms of the ratios are used to reduce the biasing effect of nonlinearity, this correlation is almost perfectly negative ($r = -.81$).

The irony of this result is that a mere increase in positive test responses may be seen as an improvement of the test when in fact its diagnostic value drops even when the test's sensitivity remains the same. This example can be extended to current research on prejudice, with the variables I and E, respectively, denoting implicit and explicit attitudes. Research with the implicit association test (IAT; Greenwald, McGhee, & Schwarz, 1998) routinely shows large effect sizes for implicit bias and

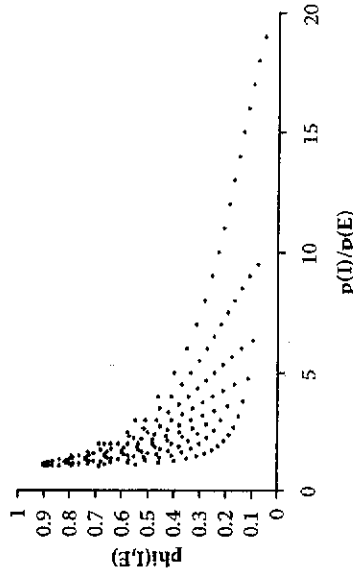


Figure 1.1 The maximum correlation between I and E plotted against the ratio of the two base rates.

low to moderate correlations between I and E. As the high values of $p(I)$ are hailed as important evidence for how many people are really prejudiced, it is overlooked that the same finding reduces the predictive power of the test. In clinical assessment the reliance on pseudo-diagnosticity is particularly fraught with danger because a person's true status is difficult to ascertain. Ultimately, criterion judgments about underlying conditions are also clinically derived. Such judgments are notoriously elastic, as they allow, for example, speculative assessments of a person's latent (i.e., not overtly expressed) pathological disposition.

To Robyn, social responsibility demands that people, and presumed experts, be disabused of associationist thinking. Ultimately, his outlook is optimistic. He defines rational thought as the avoidance of contradictory beliefs. A set of beliefs or inferences is rational if it is coherent, and to insure coherence, it is necessary to make comparisons. In Bayesian language, drawing inferences from the ratio's numerator alone is associationist and bound to be incoherent. In contrast, drawing inferences from comparisons avoids contradiction. Often, associationist thinking is considered primitive and automatic, whereas comparative thinking is controlled, effortful, and resource-consuming. Robyn generally agrees with this view, but he suggests his own characteristic metaphor. As a young man, he taught children in a summer camp to swim. He noticed that the kids instinctively tried to keep their heads above water, which moved their bodies into a vertical position, thereby making drowning more likely. To help them overcome their fear, he had to teach them to keep their faces beneath the waterline and come up for air only periodically and rhythmically. As all swimmers know, this new attitude

quickly becomes second nature, and so it can be with rational thinking, according to Robyn.

Of course, proper comparisons require the presence of all the necessary information. Sometimes people do not have access to that information but act as if they do. To illustrate, Robyn extended Tversky and Kahneman's notion of the availability heuristic to incorporate what he termed a "structural availability bias" (Dawes, 2005). Again, a clinical example is instructive. Robyn observed that many clinicians claim that "child abusers never quit on their own without therapy." All that clinicians know, however, is the probability with which child abusers stop given that they are in therapy. The probability with which abusers stop without being in therapy is out of view. When the proper input for rational judgment cannot be obtained, Robyn advised suspending judgment, an insight that is as indicative of his genius as it is counterintuitive.

Another inspiration Robyn drew from Tversky and Kahneman's work is the gambler's fallacy. In *Rational Thought in an Uncertain World* (Dawes, 1988) Robyn illustrated this bit of irrationality with a letter to Dear Abby, in which the writer insisted that the law of averages [sic] demanded that a baby boy was due after five consecutive baby girls. Whereas the gambler's fallacy is easily explained statistically and equally easily demonstrated empirically, the field of psychology has sorely been missing a compelling account of the mechanisms that produce it. Gold and Hester (chapter 2) present a series of experiments, asking just what type of animistic reasoning is involved. The answers lie in the boundary conditions of the fallacy. It disappears, for example, when a coin is exchanged for a new one before the critical trial or if a longer interval between tosses is introduced. The memory of the coin is as weak as its muscle. Suppose Abby's correspondent had *wanted* another girl. Would she have boosted her confidence in getting her wish by looking for a new father or by imposing a few months of sexual abstinence?

One important line of Robyn's work has been dedicated to attitude measurement, culminating in a book (Dawes, 1972) and in a chapter (Dawes & Smith, 1985) in the third edition of the *Handbook of Social Psychology*. This chapter contains a surprising insight about the century-old tug of war over the predictive value of social attitudes. LaPiere (1934) published a landmark study suggesting that attitudes and behaviors are unrelated. LaPiere found that a Chinese couple touring California was virtually never turned away at hotels or restaurants, yet virtually all clerks at these establishments asserted over the phone that they would refuse service to "members of the Chinese race." Dawes and Smith noted that a correlation coefficient is not defined when three of the four cells are (virtually) empty. In other words, judgments about

how well attitudes predict behavior had to be suspended in the context of LaPiere's study (which virtually no attitude researcher had done). In hindsight, LaPiere's findings can be conceptualized within the Bayesian framework discussed above. Namely, the upper bound of the attitude-behavior correlation is low inasmuch as the ratio of base rates (for a particular attitude and for a behavior corresponding to that attitude) is large.

Robyn's interest in measurement converged with his long-standing concern about overgeneralization when he turned his attention to the well-known contrast effect in attitudinal judgment. Attitudinal contrasts mean that people overestimate the difference between typical attitudes held by people with known disagreements. Judging existing methods to be unsatisfactory, Dawes, Singer, and Lemons (1972) devised a paradigm in which members of two attitudinal camps (with regard to the Vietnam War) had to write statements they considered typical of members of the opposing camp. As expected, doves wrote statements for hawks that hawks themselves rejected as too extreme and vice versa. The origins of this idea in Robyn's work on set relations are evident. Participants treated the two attitudinal sets as being exclusive, when in fact they were disjunctive.

When rational thought cannot be coaxed to the surface by able tutoring and dedicated practice, people need not give up. Recognizing how difficult it is for the human mind to avoid random judgmental errors, Robyn turned his attention to comparisons of clinical judgment with actuarial decision methods. In a survey of the literature, Meehl (1954) had found that actuarial decision making beats intuitive judgment when valid correlations between predictors and outcome's are known. Scores on predictor variables can be clerically combined, thereby guaranteeing freedom from random error and the haphazard intrusion of nondiagnostic variables. Robyn showed that unit weights, or even random weights, not only do better than holistic intuitive judgments, but that they are also more robust than optimal regression weights (Dawes, 1979; see Appendix 1, this volume). Optimal weights tend to overfit the data of the validation sample and thus lose predictive power in a new sample. To avoid the pitfalls of both holistic judgment and statistical overfitting, one only needs to "know what variables to look at and then know how to add" (Dawes & Corrigan, 1974, p. 105). In a striking demonstration, Howard and Dawes (1976) predicted couples' marital happiness from the frequency of sexual intercourse (broadly defined) and the frequency of arguments and fights. The difference between the two was a better predictor than each frequency alone because the two were positively correlated in their particular samples.

Following up on Robyn's suggestion that federal granting agencies use linear composites of specific judgments to evaluate the merit of grant proposals, Arkes (chapter 3) found that disaggregated judgments are more reliable than the holistic judgments that are usually made. Arkes's tale illuminates both how "improper linear models" can be put to good use in practice and the kind of resistance such a rational procedure may face, even among sophisticated scientists. Dana (chapter 4) presents a much needed, in-depth analysis of the conditions under which improper linear models work. He finds that such models work best when predictor-outcome associations are weak, which is to say they work particularly well in the social sciences, where the temptation to rely on intuitive appraisals is the greatest.

Faust (chapter 5) takes this approach further by applying it to judgments about the quality of scientific theories. Drawing on philosophy of science, Faust notes the existence of multiple reasonable criteria by which theories can be appraised (e.g., parsimony, falsifiability, generativity). Despite claims to the contrary, no single criterion trumps all others. Even Popper's criterion of disconfirmability is no gold standard because Popper's theory itself might be disconfirmed, and then what? Faust's theme connects with Robyn's insistence that comparisons be made (e.g., it is not enough to enumerate instances in which parsimonious theories do well); it also connects with the idea that simple combinations of criteria can go a long way toward reliable and valid appraisals. Nevertheless, Faust goes beyond the linear-combination model by suggesting that, at the present state of our knowledge, nonlinear configurations of predictor scores might yield incremental validity.

Krueger (chapter 6) addresses the issue of incremental validity in a variety of social judgment tasks. Some current models of social judgment (e.g., for self-perception or stereotyping) use predictors that empirically do not contribute independently to outcome judgments. For example, self-other comparisons are well predicted by absolute self-judgments, whereas absolute judgments of others contribute little. This finding reveals two of Robyn's themes. Namely, to be rational, people *should* base self-other comparisons on both types of absolute judgment. When they fail to do that, however, scientific modeling of the intuitive judgment process should only include the valid predictors.

The regression-analytic approach sheds light on Bayesian probability judgments. It is evident from Bayes's theorem that across multiple judgment contexts, $p(I|E)$ increases with $p(I)$ and with $p(E|I)$ and that it decreases with $p(E|-I)$. For a sample of judgment contexts, optimal regression weights can be calculated, but they are of limited use in other contexts. To obtain decontextualized, and thus nonoptimal weights,

one can begin with the assumption that the distributions of the three predictor variables are flat (i.e., one assumes uniform priors). In a simulation with predictor values ranging from .1 to .9 in steps of .2, $p(I|E)$ is correlated .81, .37, and -.37, respectively, with $p(I)$, $p(E|I)$, and $p(E|-I)$. The criterion base rate is a better predictor than the probability of the evidence conditioned on the criterion. What is more, the correlation between criterion and prediction is even larger than the correlation between $p(I|E)$ and the likelihood ratio, $p(E|I)/p(E|-I)$, which is .46. Even the natural logarithm of the ratio does not predict $p(I|E)$ as well ($r = .55$) as the base rate does.

The predictive value of Bayesian probabilities across contexts may explain why Robyn never joined the call for an abandonment of null hypothesis significance testing (NHST). In NHST, $p(E|I)$ is the p value that researchers use to reject or retain a null hypothesis. In a mature area of research, credible estimates of $p(I)$ and $p(E|-I)$ are available and should be used to estimate the probability of the hypothesis given the evidence. In many areas of "soft" psychology, however, such estimates are often lacking, and $p(E|I)$ is all the researchers have (Krueger, 2001). Thus, it is defensible to assume uniform priors for $p(I)$ and $p(E|-I)$, and to base inferences on the predictive value of $p(E|I)$. Applying his "improper" rule, Robyn noted that NHST at least tells researchers that there is "not nothing" and the directionality thereof (Dawes, 1991).

In his advocacy of rational inference, Robyn has held clinicians, academic research scientists, and ordinary people to the same standards. Putative demonstrations of biased judgment must themselves satisfy the demands of coherence. The most famous of Robyn's debunking of putative biases was his finding that the vaunted "false consensus effect" is not necessarily false. Using both a Bayesian (Dawes, 1989) and a regression-analytic approach (Dawes, 1990), he showed that people are well advised to use their own responses to predict the responses of others inasmuch as they are uncertain about what others will actually do. This heuristic reasoning guarantees that errors will be made, but it minimizes the errors relative to what would happen if people merely guessed (note the comparative nature of this argument). This pivotal insight led to a revival of research on social projection (e.g., Krueger, 1998, for a review), including an empirical article by Dawes and Mulford (1996), which also questioned whether the so-called overconfidence bias has been demonstrated satisfactorily. With characteristic candor, Dawes and Mulford (p. 201) suggested that the "belief in these particular systematic limitations of judgment arises not from the irrationality of experimental subjects who allegedly demonstrate their

existence, but from the cognitive limitations of the psychologists studying these subjects."

Moore and Small (chapter 7) open the case against the entrenched view that it is logically impossible for most people to be better than average. They show that they can, under conditions that are often met in the real social world. Using Bayesian and regression-analytic tools, Moore and Small present a theoretical model that can reproduce both better-than-average and worse-than-average effects. Their empirical data support their analysis, although they allow for motivated distortions in addition to rational inference. Bar-Hillel, Budescu, and Amar (chapter 8) pursue the better-than-average phenomenon at the group level. Noting that judgmental optimism is expressed by the overestimation of desirable outcomes, they examine whether people are more likely to predict that a sports team will win if that outcome results in their receiving a payment. In two studies they find small effects, leading them to conclude that the wishful thinking effect remains elusive.

Robyn's interest in the consensus bias originated in his work on social dilemmas. Dawes, McTavish, and Shaklee (1977; see Appendix 2, in this volume) found that cooperators in a prisoner's dilemma expected other players to cooperate, whereas defectors expected others to defect. The prisoner's dilemma is the kind of situation that maximizes a person's uncertainty regarding the behavior of others. At the limit, ignorance is complete, meaning that the probabilities of all possible rates of cooperation are uniform. Laplace's rule of succession says that after a person's own choice between cooperation and defection is in evidence, the posterior probability of the other's choice being the same is $2/3$ (see Dawes, 1989, for a derivation). Recent evidence suggests that this is why some people cooperate in the first place, namely, because they realize that making the cooperative choice enables them to expect a higher value from the game than does making a defecting choice (Krueger & Acevedo, 2007).

Robyn is not altogether comfortable with this conclusion. He argues instead that people draw on preexisting inclinations to either cooperate or to defect and then project their own choice rationally. In a series of studies conducted with his friend John Orbell at the University of Oregon, Robyn proposed and refined a "cognitive miser model," which reveals the social benefits of projection in games that allow players to withdraw (Orbell & Dawes, 1991). Because both cooperators and defectors expect others to be like them, defectors anticipate the dreary Nash equilibrium of mutual defection. When offered the option to withdraw, they take it. Cooperators also project but anticipate mutual cooperation. They are more likely to stay in the game and, hence, a greater total

amount of resources is earned than would have been if everyone had been forced to play.

In *n*-person social dilemmas, Robyn distinguished between give-some and take-some games (Dawes, 1980). Give-some games are exemplified by public goods dilemmas, such as donating to public broadcasting; take-some games have decisions to pollute as a prominent real-world analog. Empirically, many people cooperate, and the question is why. As Dawes and Messick (2000, p. 111) put it, the puzzle is that "by doing what seems individually reasonable and rational, people end up doing less well than they would have done if they had acted unreasonably or irrationally." Here is a genuine dilemma for Dawes himself, as individual rationality comes face to face with social responsibility. A single-minded pursuit of individual rationality spells disaster for the group. What to do? Merely redefining cooperation as being "collectively rational" skirts the issue (and abandons methodological individualism).

Robyn and colleagues found that communication among group members is of great value. When talking to one another, participants can construct a shared understanding of the situation they are in, form an inclusive social identity, and exchange promises of cooperation (Orbell, van de Kragt, & Dawes, 1988). These social events work. The finding that promises are honored is surprising from a narrow view of individual rationality because after listening to the promises of others, each individual could conclude that the prospect of getting a free ride is now particularly alluring.

Bicchieri (chapter 9) offers one answer for why promises work. She notes that much collectively desirable behavior depends on the activation of social norms. Making promises, reciprocating the promises of others, and honoring such promises is ethically mandated, yet social interactions remain critical. It is not sufficient to assume, Bicchieri argues, that people act in the collective interest out of social preferences (e.g., benevolence or inequality aversion). However, this is what many current reformulations of economic theories of rationality do (Fehr & Schmidt, 1999). Norms depend on the expectations individuals have about the behavior of others, whereas social preferences are decontextualized and consequentialist. People who honor norms expect others to do the same (i.e., they project) and, thus, care about *how* final collective outcomes come about.

Brewer (chapter 10) takes up the question of how cooperation can emerge even in the absence of communication. She highlights the role of social identification with a group, which experiments have shown to occur even in minimal laboratory groups (Tajfel, Billig,

Bundy, & Flament, 1971). According to self-categorization theory (Turner, Hogg, Oakes, Reicher, & Wetherell, 1987), the experience of being in a group entails a sense of depersonalization, in which individuals regard themselves as being interchangeable with other group members. It follows that, to the extent that they trust themselves, they can also trust others (and hence cooperate virtually with themselves). Brewer reviews the evidence for several mechanisms that may account for the emergence of depersonalized trust. Critically, she notes that the presence of an outgroup is not necessary. Without an outgroup, meta-contrasts between groups, as postulated by self-categorization theory, are not possible. Likewise, the comparative (here, intergroup) judgments Robyn regards critical for rational inference do not appear necessary either. Then again, Robyn also asked that only valid predictors be used in judgment.

Messick (chapter 11) focuses on the question of how cooperation can emerge and stabilize over repeated rounds of social dilemma games. When it can be shown that "good guys" can win, it follows that good groups (or firms) can win, too. Arguably, social dilemmas played out at the group or corporate level involve higher stakes for the collective good than do games among individuals, yet the dynamics are much the same from an analytical point of view. Messick's notion of "corporate social responsibility" is a direct challenge to the individual (i.e., selfish) rationality of Friedman's stripe. The outcome of Messick's review is textured. He shows that although good guys are likely to do poorly locally (i.e., are vulnerable to exploitation by neighbors), they do well globally (i.e., attract more wealth than do bad guys who are surrounded by bad neighbors). Although this finding may follow analytically from the definition of social dilemmas, Messick's data show how mutual cooperation can attain equilibrium status. Hence, his finding has important implications for the evolution of altruism for many biological and social systems.

Many of Robyn's contributions revolve around the idea that good science should be put to beneficial use. The pursuit of good evidence is not *l'art pour l'art*, but interwoven with social, clinical, and personal needs. Arguably, the social and behavioral sciences are lagging behind medical science in putting a premium on evidence-based practice. Byram, Schwartz, Woloshin, and Fischhoff (chapter 12) provide an example by applying a mental-models approach to expert and lay reasoning about breast cancer. Their methods enable them to identify shared and idiosyncratic misconceptions about risk and thereby empower ordinary people¹² to make better decisions.

Human rationality is a protean concept, with no universally accepted definition. In his early work, Robyn emphasized differences between belief and evidence, writing that a cognitive distortion occurs when an individual "maintains a belief despite exposure to information contradicting that belief" (Dawes, 1964, p. 444). He also suggested that the difference between normal and distorted reasoning may be fluid rather than categorical, so that "distortion can be viewed as the result of a normal cognitive process running amok" (p. 458). Later, he rendered rationality more restrictively in terms of coherence. In his chapter on behavioral decision making, written for the fourth edition of the *Handbook of Social Psychology*, he defined rationality "not in terms of selfishness and goals, or even in terms of the likelihood of means to achieve these goals, but in terms of avoiding outright contradictions in the policies or thought processes leading to choice" (Dawes, 1998, p. 497). The linkage between beliefs and evidence is now only indirect, but we know, Robyn argues, that if a set of beliefs is incoherent, not all of them can be true. Likewise, the notion of self-interest, although removed from the definition, remains indirectly relevant. Superficially, one could refuse to worry if one's own beliefs violate Bayes's theorem. However, as De Finetti (1931) showed, probabilistic beliefs can be translated into a series of bets. If the beliefs are incoherent, a canny opponent can make a Dutch book that guarantees that gambles will be lost.

What is the reach of rationality thus defined? From the outset, Robyn rejected the view that in a hierarchically organized mind, the lower instinctual and emotional faculties interfere with the otherwise finely working rationality of the conscious faculty (Dawes, 1964, 1976, 2001). This view has been entrenched in Western civilization since the days of Plato, Aristotle, and the Church Fathers. Instead, Robyn has argued that irrationality as a lack of coherence can arise in the conscious mind itself. The critical limitation of rational thought is that—aside from its boundedness to self-interest—it does not generate goals. Emotions, on the other hand, can signal to a person what his or her goals and values are.

Robyn was much intrigued by the case of Rudolf Höss, the commandant of the Auschwitz extermination camp. In a repugnant way, Höss exemplified rationality by acting on a coherent set of beliefs. Moreover, as Robyn noted, Höss embodied "in many ways the Platonic ideal" (Dawes, 2001, p. 38) of a man whose rational mind overruled his own emotional reactions. Like Eichmann and other Nazi leaders who could not bear the sight of individual humans being tortured, Höss felt the tug of compassion. However, he "rationally" forced himself to overcome this tug.

For the sake of social responsibility, rationality needs to be channelled by emotions rather than be cut off from them. There appear to be two modes of regulation. One mode involves respect for one's own emotional responses. When cognitively appraised, emotions are emotionally useful because they tell us "how we feel about it" (Schwarz, 2001). Robyn often refers to short stories by Bertrand Russell, in which the great philosopher argued that human compassion can overcome repressive and brutal social practices. The other mode involves respect for socially shared narratives. Like Robyn, Caporael (chapter 13) argues that humans are both reasoning and story-telling creatures. The challenge is to find a way to let the good stories win. To Robyn, the pursuit of science is an opportunity to create stories that gradually become better (see also Faust, chapter 5). It is to be hoped that scientific stories displace folk stories and mythologies with coherent stories based on solid evidence. Caporael's story about stories emerges from her core configurations model of human evolution. Indeed, she argues that stories are primary because they can never be fully displaced by evidence. Even scientific theories retain a core narrative, in the light of which evidence is sought and interpreted.

With regard to Höss, Caporael argues that he, while perhaps being individually rational, was caught up in a bad narrative, the Nazi con- strual of the world. As a counterweight, she tells the story of the Danish resistance and the rescue of the Danes not found strength and purpose in a been far less likely had the Danes not found strength and purpose in a folk tale that defined who they were and whom to include in their system of values. In Caporael's words, the Danes had a good story. Their daring success was linked with the construction of their national identity (see Brewer, chapter 10) and with the ability of individual Danes to project their acceptance of the national narrative on other Danes and, hence, trust their support. In a comment on how ethics can, at least in part, be derived from rational thought, Robyn notes that the Golden Rule works because people can infer what is pleasing and what is abhorrent to others from what is pleasing or abhorrent to themselves (Dawes, 2001).

The struggle to attain rationality in a complex, story-loving world and to square rationality with social responsibility is difficult. Robyn's quest for coherence has met with many roadblocks. His life's work—and in a small way, the chapters gathered here—is a testament to the contributions that rational thought can make to the social good, yet competing theories of rationality continue to thrive in academia, and competing narratives continue to hold sway over individual imaginations. If the past is a guide to the future, this basic reality will not change. Perhaps it is apt to conclude with a story. The members of a synagogue find that

they are split into two camps. One camp argues that a particular prayer is recited standing up, whereas the other camp argues that it is recited sitting down. To settle the dispute, they ask an elder for guidance. After listening to the case made by the first camp, he shakes his head, saying that it is not the tradition to pray standing up. Well then, they say, it must therefore be the tradition to pray sitting down. Again, the old sage shakes his head. That is not the tradition either. Exasperated, they conclude that they will have to keep arguing about this matter. *That*, the old man says, is the tradition.

This brief sketch of Robyn Dawes's views on rationality and social responsibility is necessarily incomplete. This volume ends with a list of selected publications, which the reader may consult for a deeper understanding of the breadth and richness of Robyn's contributions. Still, further reading will not fully disclose the informal impact Robyn has had on the contributors to this volume and many others. After the Festschrift symposium I asked the present authors to send one paragraph describing Robyn's personal impact on them and another paragraph describing his most significant intellectual impact. To conclude this introduction, I offer a sample of quotes—without attribution—to highlight the variety of ways in which Robyn's presence has left footprints in many careers.

Robyn does care, with a passion, not only for ideas and mathematics and music, but also about ethics, morality, and future of this species.

Robyn's work was so meaningful to me that it caused me to change careers. I was a graduate student in a clinical psychology program geared toward training practitioners. My coursework exposed me to Robyn's clinical judgment work more so than at most training programs, probably owing to the fact that a couple of years earlier Robyn had graciously agreed to enter the lion's den and speak to the program defending his *House of Cards* work. When we read "Representative Thinking in Clinical Judgment" I was amazed both at how telling were the (often unfavorable) descriptions of clinical activity and at how many clinicians either ignored or rejected the work out of hand. Eventually, I applied to be his student so that I could dedicate myself to judgment and decision making research.

Dawes and Corrigan's (1974) demonstration of the inherent limits to regression approaches for revealing the cognitive processes underlying complex judgments provided an intuitive feeling for

the seductiveness of explaining potentially random data. It was liberating research in the narrow sense of freeing me from trying to do the impossible—and in the broad sense of showing that there were insights available only to those who looked for the fundamental structure of the problem. Nobody does it like Robyn.

It was a bit of a shock to meet Robyn and find myself baffled by conversations with him. He often says things that, I must admit, I have difficulty following. His statements often seem to assume knowledge that I could not have had. When, in embarrassment, I admitted to others that I did not always understand Robyn, I was comforted to learn that others had had the same experience. The general consensus here is that there are two leading explanations of this: 1. Robyn is so much smarter than us mere mortals that it is difficult for us to follow the speed and logic of his thinking. 2. Robyn is only tenuously in touch with the basic norms of everyday social interaction. The data are equally consistent with both of these.

I feel that Robyn Dawes's most important contributions are his insistence on intellectual humility and process transparency. To make a good decision that influences the lives of other people, let people know upon what bases the decision is being made and then do it in a fair and consistent manner. People deserve no less than our best decision procedures. To do less is unethical.

As Dawes noted, in almost 2,000 experimental game studies, no one attempted to discover if cooperation would occur in the absence of incentives. . . . Dawes developed an interesting "subtractive method" for his research, where possible alternative explanations for cooperation were controlled in the experimental design, and what could not be controlled became the focus of the manipulation for the next study. Nevertheless, despite such care, there are still experts that still believe that somewhere, there must be such a hidden incentive.

I recall once at a conference in the United Kingdom, during the 1980s when I was still the editor of the *Journal of Experimental Social Psychology*, I ran into Robyn shortly after I had checked in, but before my wife and I had gone to our room. Robyn was excited about an idea that had to do, as best I could tell in the hotel lobby, with a claim that the false consensus effect was, in fact, not false at all. Robyn had a napkin or some other small piece of paper and he was making notes and writing equations on it for

me to see, and, while I grasped the general drift of his claim, I did not grasp the underlying logic. (Robyn often made me feel like a moron because I failed to see something that was blatantly evident to him.) I asked him to write out the argument and send it to me as a journal editor and I would give him some feedback on the interest of the idea. The sequel, as they say, is history. He sent me the paper; I read it and studied it and dug through a rather opaque argument (it is hard to write clearly for a moron when the validity of the argument is self-evident) and realized that his argument was original, valid, and nearly inaccessible to most of our readers. A revision or two improved the accessibility and the resulting paper is one of the most important I published as an editor. Again, Robyn Dawes was a decade ahead of his peers.

Robyn is the most interesting philosopher I know. Yes, philosopher. His understanding and critical assessment of the foundations of rationality is almost unparalleled among philosophers who do this for a living. And the most interesting thing is that he follows this up with experiments showing that what appears as irrational behavior is in fact rational, and the real problem is that we usually attribute wrong motives to agents.

REFERENCES

- Bartlett, F. C. (1932). *Remembering*. Cambridge, UK: Cambridge University Press.
- Coombs, C. H., Dawes, R. M., & Tversky, A. (1970). *Mathematical psychology: An elementary introduction*. Englewood Cliffs, NJ: Prentice-Hall.
- Dawes, R. M. (1964). Cognitive distortion. *Psychological Reports*, 14, 443-459.
- Dawes, R. M. (1966). Memory and distortion of meaningful written material. *British Journal of Psychology*, 57, 77-86.
- Dawes, R. M. (1972). *Fundamentals of attitude measurement*. New York: Wiley.
- Dawes, R. M. (1976). Shallow psychology. In J. Carroll & J. Payne (Eds.), *Cognition and social behavior* (pp. 3-12). Hillsdale, NJ: Erlbaum.
- Dawes, R. M. (1979). The robust beauty of improper linear models. *American Psychologist*, 34, 571-582.
- Dawes, R. M. (1980). Social dilemmas. *Annual Review of Psychology*, 31, 169-193.
- Dawes, R. M. (1988). *Rational choice in an uncertain world*. San Diego, CA: Harcourt, Brace, Jovanovich.
- Dawes, R. M. (1989). Statistical criteria for establishing a truly false consensus effect. *Journal of Experimental Social Psychology*, 25, 17.

- Dawes, R. M. (1990). The potential nonfalsity of the false consensus effect. In R. M. Hogarth (Ed.), *Insights in decision making: A tribute to Hillel J. Einhorn* (pp. 179-199). Chicago: University of Chicago Press.
- Dawes, R. M. (1991). Probabilistic versus causal thinking. In D. Cuihetti & W. M. Grove (Eds.), *Thinking clearly about psychology: Vol. 1. Matters of public interest. Essays in honor of Paul Everett Meehl* (pp. 235-264). Minneapolis: University of Minnesota Press.
- Dawes, R. M. (1994). *House of cards: Psychology and psychotherapy based on myth*. New York: The Free Press.
- Dawes, R. M. (1998). Behavioral decision making and judgment. In D. Gilbert, S. Fiske, & G. Lindzey (Eds.), *The handbook of social psychology* (4th ed., Vol. 2, pp. 497-548). Boston: McGraw-Hill.
- Dawes, R. M. (2001). *Everyday irrationality: How pseudoscientists, lunatics, and the rest of us systematically fail to think rationally*. Boulder, CO: Westview Press.
- Dawes, R. M. (2005). An analysis of structural availability biases, and a brief experiment. In K. Fiedler & P. Juslin (Eds.), *Information sampling and adaptive cognition* (pp. 147-152). New York: Cambridge University Press.
- Dawes, R. M., & Corrigan, B. (1974). Linear models and decision making. *Psychological Bulletin*, 81, 95-106.
- Dawes, R. M., McFavish, J., & Shaklee, H. (1977). Behavior, communication, and assumptions about other people's behavior in a commons dilemma situation. *Journal of Personality and Social Psychology*, 35, 1-11.
- Dawes, R. M., & Messick, D. M. (2000). Social dilemmas. *International Journal of Psychology*, 35, 111-116.
- Dawes, R. M., Mirels, H. I., Gold, E., & Donahue, E. (1993). Equating inverse probabilities in implicit personality judgments. *Psychological Science*, 6, 396-400.
- Dawes, R. M., & Mulford, M. (1996). The false consensus effect and overconfidence: Flaws in judgment or flaws in how we study judgment? *Organizational Behavior and Human Decision Processes*, 65, 201-211.
- Dawes, R. M., Singer, D., & Lemons, F. (1972). An experimental analysis of the contrast effect and its implications for intergroup communication and the indirect assessment of attitude. *Journal of Personality and Social Psychology*, 21, 281-295.
- Dawes, R. M., & Smith, T. E. (1985). Attitude and opinion measurement. In G. Lindzey & E. Aronson (Eds.), *The handbook of social psychology* (3rd ed., Vol. 1, pp. 509-566). New York: Random House.
- De Finetti, B. (1931). *Probabilismo: Saggio critico sulla teoria delle probabilità e sul valore della scienza*. In A. Aliviera (Ed.), *Biblioteca di Filosofia* (163-219). Naples, Italy.
- Fehr, E., & Schmidt, K. (1999). A theory of fairness, competition, and cooperation. *The Quarterly Journal of Economics*, 114, 159-181.
- Greenwald, A. G., McChes, D. E., & Schwartz, J. L. K. (1998). Measuring individual differences in implicit cognition: The implicit association test. *Journal of Personality and Social Psychology*, 74, 1464-1480.
- Hastie, R., & Dawes, R. M. (2001). *Rational choice in an uncertain world* (2nd ed.). New York: Sage.
- Howard, J. W., & Dawes, R. M. (1976). Linear prediction of marital happiness. *Personality and Social Psychology Bulletin*, 2, 478-480.
- Krueger, J. I. (1998). On the perception of social consensus. In M. P. Zanna (Ed.), *Advances in experimental social psychology* (Vol. 30, pp. 163-240). San Diego, CA: Academic Press.
- Krueger, J. I. (2001). Null hypothesis significance testing: On the survival of a flawed method. *American Psychologist*, 56, 16-26.
- Krueger, J. I., & Acevedo, M. (in press). Perceptions of self and other in the prisoner's dilemma: Outcome bias and evidential reasoning. *American Journal of Psychology*, 120, 593-618.
- LaPiere, R. T. (1934). Attitudes vs. action. *Social Forces*, 13, 230-237.
- Meehl, P. E. (1954). *Clinical versus statistical prediction: A theoretical analysis and a review of the evidence*. Minneapolis: University of Minnesota Press.
- Meehl, P. E., & Rosen, A. (1955). Antecedent probability and the efficiency of psychometric signs, patterns, and cutting scores. *Psychological Bulletin*, 52, 194-216.
- Orbell, J. M., & Dawes, R. M. (1991). A "cognitive miser" theory of cooperators' advantage. *American Political Science Review*, 85, 515-528.
- Orbell, J. M., van de Kragt, A. J. C., & Dawes, R. M. (1988). Explaining discussion-induced cooperation. *Journal of Personality and Social Psychology*, 54, 811-819.
- Schwarz, N. (2001). Feelings as information: Implications for affective influences on information processing. In L. L. Martin & G. L. Clore (Eds.), *Theories of mood and cognition: A user's guidebook* (pp. 159-176). Mahwah, NJ: Erlbaum.
- Tajfel, H., Billig, M. G., Bundy, R. P., & Flament, C. (1971). Social categorization and intergroup behavior. *European Journal of Social Psychology*, 1, 1-39.
- Turner, J. C., Hogg, M. A., Oakes, P. J., Reicher, S. D., & Wetherell, M. (1987). *Rediscovering the social group: A self-categorization theory*. Oxford, UK: Blackwell.
- Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, 185, 1124-1131.

