Mem Cogn (2017) 45:566–575
DOI 10.3758/s13421-016-0686-8

A meta-analysis of response-time tests of the sequential two-systems model of moral judgment

Jonathan Baron[1] · Burcu Gürçay[1]

Published online: 27 December 2016
© Psychonomic Society, Inc. 2016

Abstract The (generalized) sequential two-system ("default interventionist") model of utilitarian moral judgment predicts that utilitarian responses often arise from a system-two correction of system-one deontological intuitions. Response-time (RT) results that seem to support this model are usually explained by the fact that low-probability responses have longer RTs. Following earlier results, we predicted response probability from each subject's tendency to make utilitarian responses (A, "Ability") and each dilemma's tendency to elicit deontological responses (D, "Difficulty"), estimated from a Rasch model. At the point where A = D, the two responses are equally likely, so probability effects cannot account for any RT differences between them. The sequential two-system model still predicts that many of the utilitarian responses made at this point will result from system-two corrections of system-one intuitions, hence should take longer. However, when A = D, RT for the two responses was the same, contradicting the sequential model. Here we report a meta-analysis of 26 data sets, which replicated the earlier results of no RT difference overall at the point where A = D. The data sets used three different kinds of moral judgment items, and the RT equality at the point where A = D held for all three. In addition, we found that RT increased with A–D. This result holds for subjects (characterized by Ability) but not for items (characterized by Difficulty). We explain the main features of this unanticipated effect, and of the main results, with a drift-diffusion model.

Keywords Moral judgment · Dual-systems · Response Time · drift diffusion

Introduction

Recent research on moral judgment has focused on the conflict that sometimes exists between judgments consistent with utilitarianism and those consistent with deontological rules or emotional responses. A typical dilemma involving such conflict concerns whether it is right to kill one person in order to save five others. Utilitarianism implies that the better of two options is the one that leads to the better overall outcome for everyone affected (fewer deaths, in this case).[1] By contrast, deontology consists of an alternative group of moral theories based on rules, which can define right and wrong action independently of its consequences, e.g., "Don't kill people." In using the terms "utilitarian" and "deontological", we do not imply that our subjects fully accept either moral theory, only that their judgments are consistent with one theory or the other in the dilemmas of interest. As we note later, individuals differ substantially in whether their judgments tend to follow utilitarianism or not.

The substance of this article (with 24 studies instead of 26) was presented at the meeting of the Society for Judgment and Decision Making in Chicago, November 2015. We thank Ian Krajbich for helpful comments on that presentation, and Denise Cummins for useful discussions.

✉ Jonathan Baron
baron@psych.upenn.edu

[1] Department of Psychology, University of Pennsylvania, 3720 Walnut St., Philadelphia, PA 19104–6241, USA

[1] Utilitarianism, unlike some other theories in which moral conclusions are inferred from consequences and their probabilities, is sensitive to the number of people affected in different ways. Many of the scenarios used in this research involve numbers, so the often-used term "consequentialism" does not quite get at the issue.

A prominent account of how people resolve this kind of conflict is what we call the sequential two-systems model of moral judgment (Greene, Sommerville, Nystrom, Darley, & Cohen, 2001). This model proposes that moral dilemmas vary systematically in terms of how much the deontological response is supported by intuition or immediate emotion. When the deontological response ("don't kill") comes immediately to mind, often because it is evoked by some emotion, it sometimes determines the subject's response. At other times, reasoning acts to correct an intuitive response. The correction leads to a utilitarian response. The idea that an intuitive response is sometimes corrected by a more reflective system is part of a more general theory advocated by others such as Kahneman (2011) and Evans (2007; Evans & Stanovich, 2013), who call the same idea "default interventionst." These are not the only form of two-systems theory (Sloman 1996), but the sequential assumption is the one at issue here.

This assumption is supported by various kinds of evidence, on which we comment later: response times to utilitarian responses are longer; cognitive interference and time pressure sometimes have a greater effect on utilitarian responses; and instructions to respond intuitively or reflectively affect the relative probability of the two response types.[2] Gürçay and Baron (2017) review this evidence and present results that support an alternative model. One of these results, a conceptual replication of Koop (2013), involves tracking the mouse pointer while people make judgments under conditions of conflict. The sequential model predicts that subjects would occasionally move the mouse toward the intuitive/deontological response and then change their minds and move toward the reflective/utilitarian alternative. This should happen more often than a switch in the opposite direction. Both Gürçay and Baron (2017) and Koop (2013) found no general difference in the number of switches in the two directions. These results suggest that conflict can be resolved either way and that the two different principles — deontological rules vs. total consequences — vie with each other from the outset. That is the alternative model that we examine here, which we call the conflict model.

Some of the evidence for the sequential model is that response times (RTs) are longer for utilitarian responses. Baron, Gürçay, Moore, and Starcke (2012) pointed out that RT is usually longer when response probability is lower, other things being equal. Thus a more important test involves equating response probability. Our conflict model predicts that response time (RT) to utilitarian and deontological responses should be the same when the probabilities of the two responses are equal at .5. The sequential model predicts that, even here, utilitarian responses should take longer on the average because some of these responses result from reflective correction of intuitive deontological responses. This issue is the main topic we address here.

The sequential model as stated so far implies that the reflective/utilitarian process does not begin until the intuitive/deontological process is complete, as a kind of checking step. Note that other versions of the two-system theory make the same prediction. Greene in fact does not specifically state that the reflective process must begin only after the intuitive process is complete.[3] In particular, consider a version in which an intuitive/deontological process starts out immediately when the dilemma is understood and approaches an asymptote of strength fairly soon, and a reflective/utilitarian process also starts out immediately but takes some time to build up its strength, ultimately reaching full strength after the intuitive process is largely "finished", if the response has not already occurred. Because the intuitive process varies in its asymptote and the speed of approaching it, its strength will not always be sufficient to produce the response, thus allowing time for the reflective process to build up strength and control the response. Yet, when the intuitive process controls the response, RT will be relatively low, even if this happens with probability .5 for a given dilemma.

This situation is illustrated in Fig. 1. Here the three curves represent the build-up of strength in the intuitive process, the reflective process, and their difference. It is the difference that determines the response. The upper and lower dashed lines indicate the thresholds imposed on the difference. The curves present the average value, but trial-to-trial variance is considerable. So this variance will sometimes cause the intuitive process to hit its boundary, and the intuitive response will be made. This is more likely to happen when the average of the curve is closer to its boundary. Clearly this would happen sooner for intuitive response than for the reflective response. This generalized form of the sequential model thus makes the same prediction as the simple form: when response probabilities are expected to be equal, the reflective response will take longer.

Baron et al. (2012) analyzed data from 5 studies, each of which involved giving several dilemmas to several subjects. Each subject responded only once to each dilemma, so we could not find any particular dilemma that had a

[2] Evidence from neuroscience is also presented in support of the two-system model. We do not discuss this evidence because it seems to rely on a longer chain of assumptions, and we think that a cognitive model ought to have psychological implications that can be observed directly.

[3] "According to the view I have sketched, people tend to have emotional responses to personal moral violations, responses that incline them to judge against performing those actions. That means that someone who judges a personal moral violation to be appropriate (e.g., someone who says it's okay to push the man off the bridge in the footbridge case [where this will prevent a runaway trolley from hitting 5 people]) will most likely have to override an emotional response in order to do it. This overriding process will take time, and thus we would expect that 'yes' answers will take longer than 'no' answers" (Greene 2007, p. 44).

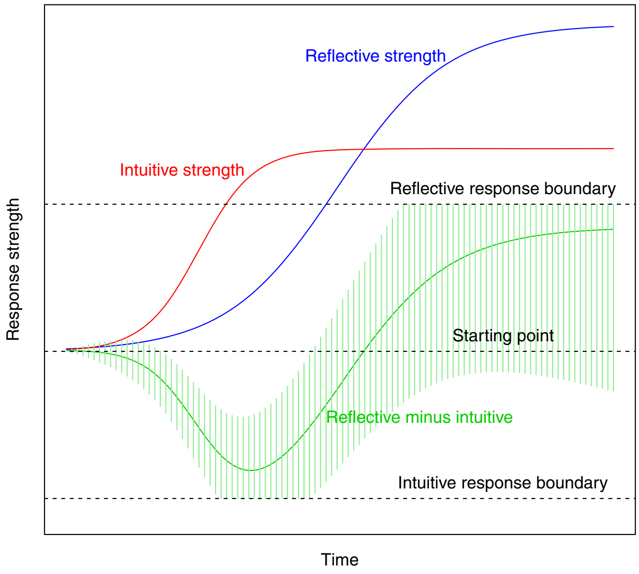
Reflective strength
Intuitive strength
Reflective response boundary
Starting point
Reflective minus intuitive
Intuitive response boundary
Response strength
Time

Fig. 1 Schematic diagram of the generalized sequential model in which the intuitive process builds strength faster

50 % probability of a utilitarian response for that subject. Instead, we attempted to determine how RT for the two possible responses (Yes for utilitarian, No for deontological) changes as a function of response probability and then estimate what would happen if probability were .5. To estimate RT at this point, we fit a Rasch model to the Yes/No data, of the sort used in educational testing.[4] We assumed (arbitrarily) that the utilitarian response was the correct answer. The model assigns to each subject an "Ability" to produce correct answers, and to each dilemma a "Difficulty". Response probability is assumed to be a logistic function of Ability-minus-Difficulty. Thus, the probability of a utilitarian response is .5 when Abiltiy equals Difficulty.[5]

The function of the Rasch model may be clarified by comparing it to an alternative approach, in which we define ability as the proportion of utilitarian responses made by each subject and difficulty as the proportion of deontological responses made to each dilemma. The Rasch model is an improvement over this approach for two reasons. First, it is more responsive to differences in proportions near 1 or 0, which may indicate substantial differences in true ability or difficulty. Second, and more importantly, it is designed to ensure that the response probability is .5 when ability and difficulty are equal. If a subject has a .5 probability of a utilitarian response, and a given dilemma gets utilitarian responses half the time across subjects, we still do not know that the subject's probability on this particular dilemma is .5. The .5 proportion for the dilemma depends on what the other subjects do. If they are mostly deontological, with proportions below .5, this dilemma could be very "easy" for the subject in question. The model scales ability and difficulty so as to get the best fit, assuming that the response probability is in fact .5 when ability and difficulty are the same. (The assumptions behind this simple version of the Rasch model are surely false, but the departures from assumptions are likely to be greatest in the extremes — e.g., asymptotes not equal to 1 and 0 — not at the middle point of greatest interest — see Baron et al., 2012.)

At this point of equality, the sequential model predicts that utilitarian responses will still take more time on the average. Our conflict model predicts that RT will be the same at this point. We estimate RT at this point by, in effect, fitting straight lines to two relationships: RT as a function of Ability−Difficulty for utilitarian responses and the same thing for deontological responses. We assume (as is verified in the data) that RT increases as probability of the response decreases. Figure 2 shows what these two lines might look like.

What we actually do is equivalent to fitting two lines but is computationally simpler. We fit the following model to the combined data from all subjects:

$$RT = b_0 + b_1 AD + b_2 U + b_3 AD \cdot U \qquad (1)$$

Here, A is the subject's ability and D is the dilemma's difficulty. RT is log(Response time). U is the response: 1 for the utilitarian option (which we call Yes), −1 for the intuitive option (No). AD is Ability−Difficulty. When AD is 0 RT for yes and no responses should be equal. So the main effect of U, the intercept at this point, is the critical

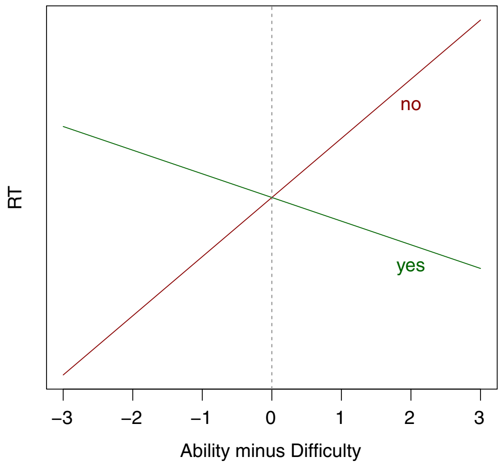

Fig. 2 Schematic diagram of the model that we fit to the data from each study

[4] We used the ltm package from R for this (Rizopoulos 2006), and, when this did not converge, we used glmer() from the lme4 package (Bates, Maechler, Bolker, & Walker, 2015).

[5] As Ability increases (because the subject is more utilitarian) or Difficulty decreases (because the deontological option seems weaker to everyone, even if they still choose it), the probability of a Yes response increases and ultimately approaches an asymptote of 1.0. The opposite happens when Ability decreases or Difficulty increases.

parameter, the distance between the two lines at this point. AD also tells us about the relative speed of the two possible responses. The interaction term $AD \cdot U$ shows how RT to Yes and No responses changes as a function of response probability, i.e., as a function of AD. Yes gets faster and No gets slower as AD increases. We fit this model using the lmer() function in the lme4 package of R (Bates et al. 2015), with crossed random effects for subjects and dilemmas, and random slopes for response U (Yes/No, coded as $1/-1$) within both.

Figure 3 shows a summary of the data from one experiment (chosen because it illustrates the points we want to make). The sizes of the circles show the effect of Ability−Difficulty on probability of responses. Happily for us, the prediction of responding is not deterministic, so subjects make enough deontological responses when Ability−Difficulty is positive, and enough utilitarian responses when it is negative, so that we can estimate the difference at zero. Note that there are very few of these "unpredicted" responses when the difference is far from zero, so that these responses have little effect on the model estimates.

The present paper extends the analysis done by Baron et al. (2012) in two ways. First, it extends the data from 5 studies to 26. This permits a meta-analysis of the main effect of interest, the value of the U parameter, which represents the difference at $AD = 0$. Although Baron et al. found that most studies showed no difference here, one of them did show the difference predicted by the sequential model. We find here that this experiment was not typical, and that some show the reverse difference.

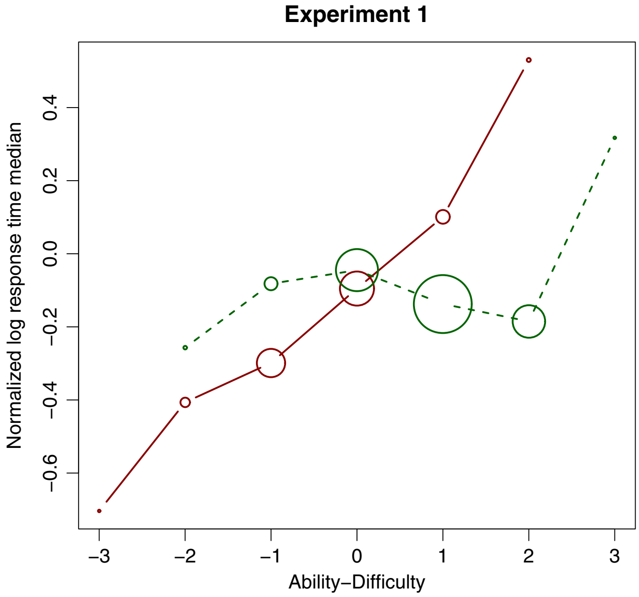

Fig. 3 Illustration of fit of response-time predictions to data from the one experiment (labeled BG2). The *red line* is for deontological responses, green for utilitarian. The areas of the *circles* are proportional to the number of observations for each point. We use across-subject medians, and rounding of Ability−Difficulty, for clarity of presentation

Second, the possibility of a meta-analysis allows us to examine the other parameters aside from U. The interaction term $AD \cdot U$ tests (and confirms) our assumption about RT being responsive to factors that affect response probability. More interesting is the AD parameter, which, on the basis of our earlier studies, we could say little about. It turns out that this is consistently positive. That is, RT increases with Ability−Difficulty. Utilitarian responses are thus slower overall, even though they are not slower at the point predicted by the sequential model. Further exploration of this difference indicates that it is probably the result of individual differences among subjects, with "higher ability" (i.e., more utilitarian) subjects taking longer.

Note that none of the data sets we use was collected to test either of the hypotheses that we consider. Thus, we do not need to worry about selective publication of positive results.

In order to understand our results better, particularly the last result, we asked whether they were consistent with a simple drift-diffusion model (of the sort first described by Ratcliff, 1978). Such models are like that in Fig. 1. They suppose a somewhat random process of drift, with the response emitted when the drift reach one boundary or the other. In the simple form of the model, the drift is linear (with noise), at some average rate over time, so that we do not expect any of the sort of curving shown in Fig. 1. The response time and probability depend on the position of the starting point relative to the two boundaries, and the expected drift rate (which may be positive or negative) relative to the boundary separation. The starting point and drift rate may depend on such factors as dilemmas and subjects. A special feature of this particular type of model is that it explains the difference between the RTs of expected and unexpected responses (e.g., a deontological response when Ability is high or Difficulty is low). Unexpected responses are often slower, as they are here.

Method

We tested the two-system model on data sets from 26 experiments (Table 1). We included the data from the three experiments in Gürçay and Baron (2017) plus the pilot study mentioned in that paper, Moore and Starcke data from Baron et al. (2012), Experiments 1–4 from Baron, Scott, Fincher, and Metz (2015) (BSFM in Table 1), Experiments 1–4 from Baron, Gürçay, and Luce (BLG, in press), and the data from Suter and Hertwig (2011), Koop (2013, both experiments), Cummins and Cummins (2012, Experiment 1), and Conway and Gawronski (2013, Experiments 1 and 2).

Table 1 Results of fitting the RT model based on Rasch scores, as described in the text: U (utilitarian responding coded as 1/0), AD (Ability−Difficulty), and AD*U (the interaction) are unstandardized regression coefficients

Study	U	(s.e.)	AD	(s.e.)	AD*U	(s.e.)
Exp 4 GB	0.082	0.023	0.039	0.056	−0.097	0.057
Moore 2011	0.054	0.027	−0.006	0.022	−0.157	0.017
Exp 3 BSMF	0.044	0.052	0.071	0.059	−0.089	0.039
Exp 2 BSMF number	0.043	0.053	0.077	0.048	−0.058	0.026
Exp 4 BLG number	0.038	0.040	0.125	0.077	−0.105	0.042
Koop1	0.035	0.024	0.021	0.015	−0.049	0.014
Exp 1 BLG rule	0.029	0.029	−0.067	0.053	−0.001	0.024
Exp 1 BSMF	0.024	0.037	0.052	0.041	−0.131	0.008
Exp 1 Conway	0.018	0.090	0.005	0.052	−0.147	0.043
Moore 2008	0.018	0.024	−0.006	0.025	−0.128	0.028
Cummins Exp1	0.011	0.029	−0.005	0.014	−0.002	0.017
Starcke 2011	0.007	0.022	0.004	0.021	−0.083	0.012
Exp 1 GB	0.006	0.033	0.031	0.026	−0.123	0.021
Exp 4 BLG rule	−0.001	0.039	0.006	0.063	−0.028	0.032
Exp 4 BSMF number	−0.005	0.059	−0.017	0.043	−0.089	0.039
Exp 2 BSMF rule	−0.022	0.032	0.011	0.037	−0.034	0.019
Exp 2 BLG rule	−0.023	0.043	0.015	0.055	−0.040	0.040
Exp 1 BLG number	−0.024	0.036	0.052	0.036	−0.056	0.032
Exp 4 BSMF rule	−0.026	0.056	−0.017	0.043	−0.089	0.039
Exp 2 GB	−0.037	0.030	−0.029	0.027	−0.121	0.024
Exp 2 BLG number	−0.051	0.037	0.079	0.039	−0.031	0.030
Koop2	−0.052	0.022	0.040	0.020	−0.086	0.013
Suter/Hertwig	−0.053	0.036	0.022	0.029	−0.010	0.184
Pilot, GB	−0.053	0.025	0.018	0.030	−0.078	0.026
Exp 3 BLG	−0.059	0.050	0.194	0.063	−0.035	0.031
Exp 2 Conway	−0.069	0.137	0.011	0.044	−0.056	0.050

The data sets involved three different kinds of dilemmas. When an experiment included both types, we analyzed them separately. Most studies had about 10 dilemmas in each condition analyzed.

The experiments labeled "Greene" in Table 1 are derived from the set used by Greene et al. (2001), as developed and modified by Greene and others. These are sacrificial dilemmas in which one person must be killed in some fairly gruesome way in order to save the lives of several others. For example, one involves dropping a cinder block on the head of a worker, killing him but preventing him from accidentally detonating an explosive that will cause a building to collapse and kill many others. The situations described are improbable at best, and that fact may make them unrepresentative of more realistic situations.

A second type, called "Ritov" (because they were first used by Ritov & Baron, 1999) had the same structure, harm caused to one person (or species, etc.) in order to prevent similar harms to many others, but they were designed to be more realistic, often based on real cases, such as setting small fires in a redwood forest in order to prevent a larger fire from spreading. Both the Ritov and Green types confounded action/omission with utilitarian/deontological. The act was always the option that caused harm to the one in order to prevent harm to the many.

A third type, called "Rule", was developed by Baron et al. (in press) with the idea of pitting a utilitarian option against an option to follow a specific deontological rule. For example, should a person lie under oath if this is the only way to insure that an innocent person is not convicted of a crime? Both options involve some action, e.g., lying vs. telling the truth. And the utilitarian response is made attractive by its beneficial effects that do not involve numbers of people affected. Baron et al. (in press) found that utilitarian responses to these three types of items correlated with each other. This result suggests that people differ in their willingness to choose utilitarian responses and not just in their willingness to act vs. do nothing.

To test the models on a set of data, we computed for each subject and each dilemma the Ability−Difficulty measure from the Rasch model, and we called this AD. We defined the response to the dilemma as U, which was 1 for a utilitarian response and −1 for a deontological response. Then we fit a model to the data set in which we predicted log RT from AD, U, and AD*U, the interaction of AD and U. We included random effect terms for subject and dilemma, and random slopes for U with respect to both subject and dilemma, using the lmer() function of the lme4 package of R (Bates et al. 2015).

We are interested mainly in the coefficient for U. Note that, because of the way we have coded U, this coefficient should represent the difference between the two best-fitting lines, one for each response type, like the lines shown in Fig. 3 except that these are straight, at the point where AD is 0. This difference should be 0 according to the conflict model but positive according to the sequential two-system model. The interaction term in the model represents the differences between the slopes of the lines, and it should be negative given our coding of the variables. The coefficient for AD represents the difference in the absolute value of the two slopes; it is positive if RT increases with AD. Note that such an increase can occur even if the U coefficient is 0.

The Suter and Hertwig (2011) data used here are based on 8 dilemmas, excluding "wallet" and "donate" (which do not involve conflicts between deontology and utilitarianism). The "Conway" data are from Conway and Gawronski (2013). This study included 10 sacrificial dilemmas like those labeled Greene, and 10 additional dilemmas in which the harm prevented by action (e.g., killing one person) was reduced substantially so that, in our terms, the dilemmas were much more difficult. Despite this difficulty, some subjects favored action in many of them, a response inconsistent with both of the usual utilitarian and deontological approaches. We eliminated subjects who favored action in 5 or more of these 10 additional dilemmas, leaving 90 out of the original 110 subjects from Study 1 and 47 out of 58 for Study 2, and then we used only the 10 sacrificial dilemmas.

The meta-analysis was done with the rma() function of the metafor package for R (Viechtbauer 2010) and plotted with the forest() function. We used the default settings, which treat the studies as a random effect, i.e., a sample from a potentially larger population of similar studies.

Results

Table 1 shows the coefficients for all data sets. Figure 4 plots the meta-analysis. The diamond in the bottom shows the mean and 95 % confidence interval for the summary of all studies. It is apparent that there is essentially no difference in RT when response probability is expected to be .5. The U coefficient mean was .0008 (not standardized, $p = 0.926$, 95 % confidence interval −0.017 to 0.019). The studies, however, do differ (with significant heterogeneity, $p = .017$), but this heterogeneity disappears ($p = .216$) with the removal of a single outlier (Exp 4 GB, at the top of Fig. 4).[6]

The possibility of a meta-analysis allowed us to examine the other parameters of the model, and these are shown in Figs. 5 and 6 for the AD and $AD \cdot U$ parameters. The latter shows the expected interaction, which is large and highly significant. More interestingly, Fig. 5 shows a consistent, if small, effect, in which — beyond the two other effects just noted — RT is slower when Ability−Difficulty is greater, i.e., when utilitarian responses are to be expected and are more often made (mean 0.015, $p = 0.0073$, 95 % c.i. 0.004 to 0.026). This general tendency for RT to increase as the probability of utilitarian responses increased is shown in Figs. 2 and 3.

The models we have fit include variance due to both subjects and items. To test whether this slowing with Ability−Difficulty is the result of item differences, subject differences, or both, we fit a simpler model to the data from each study. For each study, we regressed log RT on Ability and Difficulty, ignoring their interaction (and using ordinary regression, with unstandardized coefficients). Ability is a property of subjects and Difficulty is a property of items. We can then ask about the effect of each. Put simply, the effect of Ability−Difficulty was due to Ability, not Difficulty. Subjects whose responses were more utilitarian took more time, but "easier" scenarios did not lead to longer response times.

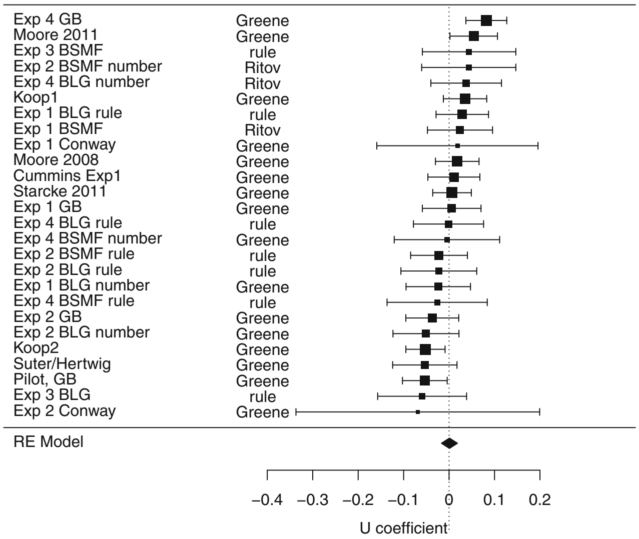

Fig. 4 Meta analysis of the results for the U parameter (RT differences between utilitarian and deontological responses when response probability is projected to be .5), also showing the type of dilemma used in each study

In a meta-analysis of these two effects across the 26 studies, the estimated effect of Ability was 0.065 ($p = .029$) and that of Difficulty was 0.006 ($p = .396$), slightly opposite to the predicted direction of this effect.[7]

Although the sequential two-system model could explain a result like this in terms of some subjects making more use of system-2, it cannot explain the present result, because it also predicts that utilitarian responses should take longer when Ability−Difficulty is zero.

The result in question is that RT for both choices is longer when utilitarian responses are more likely (positive Ability−Difficulty) but shorter when they are less likely (negative Ability−Difficulty), given the fact that RTs for the two responses are equal when they are equally likely. This result appears to be due to individual differences among subjects: subjects who make more utilitarian responses (higher "Ability") are slower. The result is consistent with other results showing that utilitarian responding is correlated with reflection (in contrast to impulsivity), the tendency to spend more time in problem solving tasks and make fewer errors (Baron et al. 2015).

To attempt to explain all the results, we asked whether a simple drift-diffusion model could account for them in principle. This model assumes that RT and choice are determined by a drift process, where some sort of information is accumulated over time, with a random element, like a

[6] This study, the discrepant study in Baron et al. (2012), involved mouse tracing, but that by itself cannot account for the unusual result, as the two Koop studies also used this procedure.

[7] T-tests across the 26 studies showed almost the same results, a significant effect of Ability (.069, $t_{25} = 2.32$, $p = 0.029$) and no effect of Difficulty (.003, $t_{25} = 0.37$, $p = 0.718$).

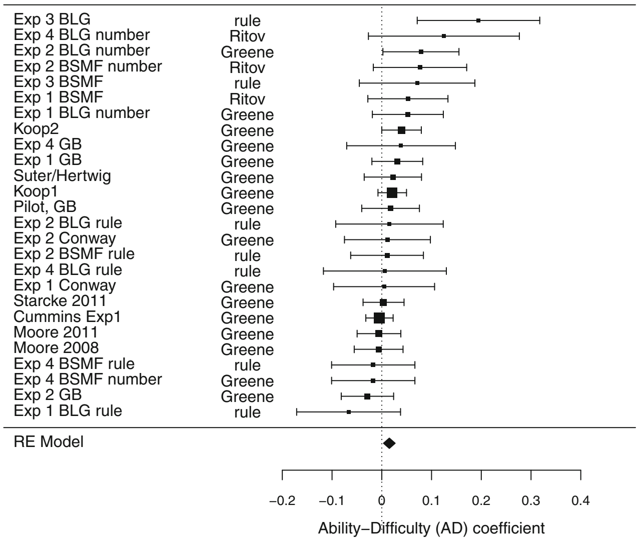

Fig. 5 Meta analysis of the results for the AD parameter (RT as a function of Ability−Difficulty)

drunk walking down a football field. When the drift process reaches one of the boundaries, the response is made. The average drift rate v may tend toward one boundary or the other, or it may be zero, in which case only the random processes can ever yield a response. The separation of the boundaries a may vary. The starting point z, relative to the two boundaries, may also vary, creating a bias toward one response or the other. In addition, the drift rate v may vary from trial to trial, with a standard deviation of sv, an additional source of randomness. (There are other parameters, but we do not need them here.) Importantly, such a model captures the spirit of the conflict model that we favor, because the drift rate does not change (except by random noise) over the course of a single choice trial, in contrast to the model illustrated in Fig. 1.

We did not attempt to fit the individual data, the data from individual studies, or the entire set of combined data. The data from individual subjects or studies seemed insufficient, and the entire data set contained large variation resulting from methodological differences across studies. Instead, we asked whether a diffusion model could account for the major features of the data that we observed. Although we could see these by looking at plots of each study, such as Fig. 3, these plots were generally quite noisy, even with the aggregation that we did in making them. So, to get another look at the overall features of the data, we plotted a diagram for all the data combined. This is Fig. 7. To make the figure, we standardized the log RTs from each study before combining the studies. In the figure, each point represents different proportions of responses from each study. However, the overall shape is consistent with our impression of the individual plots (five of which are shown in Baron et al., 2012).

Figure 7 shows four features of the data. First, the curves for the two responses cross, as noted. Second, RT is higher as AD (Ability−Difficulty) increases, so that the line for deontological responses has a steep positive slope and the line for utilitarian responses has a less steep negative slope. Third, the lines are close to linear in the region where responses are frequent. (This feature makes it reasonable to fit the linear model, shown in Fig. 2, which we used to estimate the parameters for each study.) Fourth, the linearity

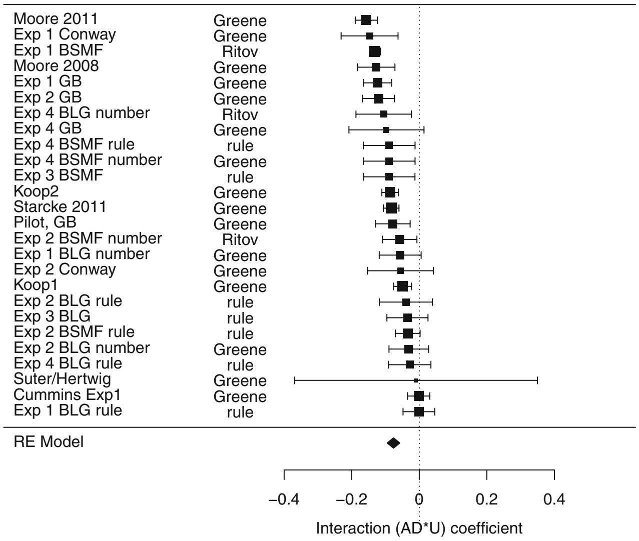

Fig. 6 Meta analysis of the results for the $AD \cdot U$ parameter (the interaction between Ability−Difficulty and type of response, utilitarian or deontological)

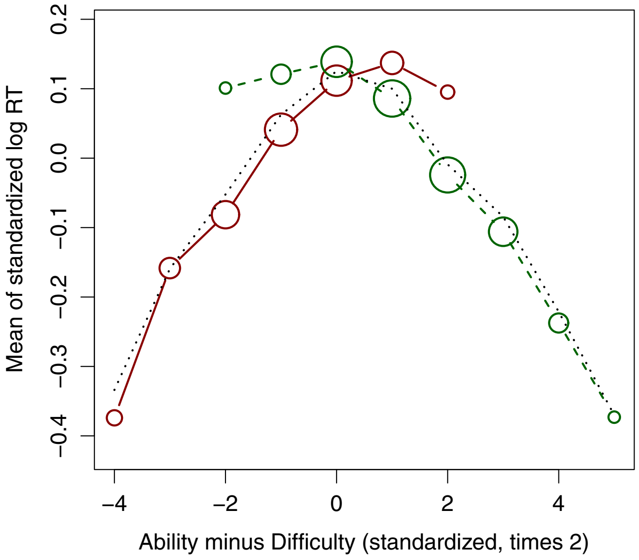

Fig. 7 RT as a function of Ability−Difficulty and type of response, utilitarian or deontological, combining data from all studies. *Red, solid, line* represents deontological responses; *green, dashed*, represents utilitarian responses. Areas of the *circles* are proportional to the number of responses. The *dotted black line* is the log mean RT collapsed over response types

does not extend through the midpoint; RT is faster on the side of unexpected responses than would be predicted by extrapolation of a straight line of the sort shown in Fig. 2. At the midpoint, response conflict is greatest. This is also apparent in the overall mean RT (dotted line).

To see if we could generate a pattern with these four features, we used the rdiffusion() function in the rtdists package of R (Brown, Gretton, Heathcote & Singmann, 2016), which simulates, with draws of random numbers, the simple diffusion model described above. We tried several combinations of parameters, and we present the best simulation in Fig. 8.[8] The drift rate v is 0 when AD is 0, and it varies linearly with AD, so v is negative when a deontological response is expected. (The assumption of linearity with AD is clearly an approximation.) In order to get the cross-over (the interaction) we needed to add some variability from trial to trial ($sv = 1$) in the drift rate v. Without this variability, the two curves lie on top of each other.[9] The model assumes that drift rate v and boundary separation a both increase with AD (Ability−Difficulty). The change in boundary separation is what produces the overall increase in RT with AD. People who are more prone to making utilitarian responses require more information before making any response.

The model, shown in Fig. 8, captures four features well, and the predictions of RT and accuracy are not too bad, despite the crudeness of the Rasch model itself, and despite the fact that the aggregation of all data by z scores of logs requires several assumptions that are surely only crude approximations (e.g., that error is the same across subjects).

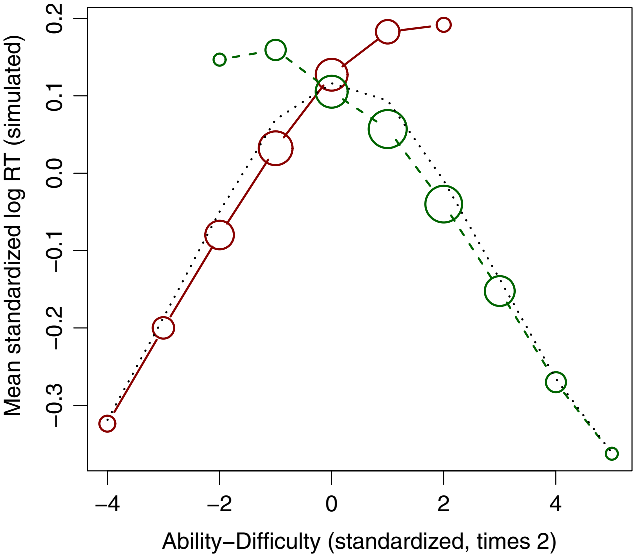

Fig. 8 Diffusion model simulation in which drift rate and boundary separation both vary with Ability−Difficulty. The *red, solid, line* represents deontological responses; *green, dashed*, represents utilitarian responses. Areas of the *circles* are proportional to respons probability. The *dotted line* is the log mean RT collapsed over response types

It may be of interest that AD appears to affect the drift rate v but not the starting point z; the model fits no better when the starting point z is allowed to depend on AD along with the drift rate. Psychologically, the drift rate v corresponds to the relative rates of accumulation of evidence for the two responses, while the starting point z corresponds to response bias. Bias could vary across subjects but apparently does not vary enough to affect the overall results; we would have to to fit models to individual subjects in order to tell, and we felt that the data were insufficient for this.

Discussion

We suggest that the sequential two-system model might work for a few dilemmas, but we think its use is not a robust result, and it may have to do with properties of the dilemmas other than the fact that they pit utilitarian vs. deontological responses, or even the fact that they put emotion on the deontological side. For example, it might have to do with the fact that they use numbers, and the utilitarian response requires attention to the numbers, which may add to the RT when subjects attend to them. Or, in other cases, subjects might attend to the numbers first and take more time when they process the additional information about the means used to bring about the outcomes. In general, though, the meta-analysis indicates that, for all the studies we could find, the overall difference predicted by the sequential model is not there. This conclusion agrees with the conclusions of other studies (Baron et al. 2012; Gürçay and Baron 2017; Koop 2013).

[8] These were optimized by hand, cycling through the relevant parameters and adjusting them in steps that were appropriate for each one, assessing the overall fit to the mean RTs and response probabilities used to construct Fig. 7. In the result shown, the parameters were as follows: boundary separation a was $2 + step/40$ where $step$ ranges from 1 to 10 for the 10 points, and the divisor 40 gave a better fit than 30 or 50; drift rate v was $(step − 5)/1.4$, and 1.4 was better than 1.3 or 1.5; the starting point z was $a/2$ (the middle), and letting it vary as a function of $step$ did not help at all; drift-rate variability sv was 1.0, which was better than 0.9 or 1.1; starting-point variability sz was 0, and increasing it did not help at all; and the time constant $t0$ was .38, which was better than .37 or .39, but which has no psychological interpretation because of the standardization of RT across subjects.

[9] This is because the RT distributions for the two responses must be identical except for a constant ratio at each time point. To see this intuitively, suppose that each step must be up or down, with respective probabilities U and D (where $U + D = 1$). And it takes S steps to reach a boundary. The the probability of reaching the top boundary in exactly S steps is U^S, and likewise D^S for the bottom boundary, so the ratio is $\frac{U^S}{D^S}$. To reach either boundary in $S + 2$ steps — an odd number of additional steps won't work — we multiply each of these by UD (the probability of the two added steps occurring, with the down step at one position in the sequence) times the number of positions in the sequence where added down steps can fall. The ratio stays the same; it is always $\frac{U^S}{D^S}$, since we have just multiplied both sides by a constant, the probability of an up and a down times the number of ways the additional down steps can occur, which is independent of U and D. Likewise for $S + 4$, and so on.

If the sequential model, even in the modified form shown in Fig. 1, cannot account for the RT data, how could we explain the other results that seem to support it? As discussed by Gürçay and Baron (2017) and by Baron (in press), the major results concern efforts to manipulate reflection vs. intuition. Utilitarian responding is sometimes reduced by cognitive load (e.g., Trémolière, De Neys, & Bonnefon, 2012; Trémolière and Bonnefon, 2014; Conway and Gawronski, 2013), and it may also be affected by other manipulations such as reading in a foreign language (Costa et al. 2014) or other possible manipulations such as time pressure (Suter and Hertwig 2011) or instructions to reflect.

Although these results are consistent with some sort of two-system model, they do not uniquely support the sequential version of it. Even if two processes start at the same time and run in parallel, one of them might require more resources than the other. Moreover, in all of these studies that we know of, researchers do not report tests of the generality of their effects across dilemmas. When these tests are done, the results are either not conventionally significant or else the number of dilemmas is too small to tell. Gürçay and Baron (2017) provide additional details, as well as reporting some failures to find effects of some of these manipulations (using tests that take scenario variance into account).

In particular, it appears from raw data we have seen that different dilemmas respond differently to cognitive load, even in opposite directions. It seems at least possible that the effect of these manipulations is to reduce, or increase, attention to some particular aspect of a dilemma, which may vary from case to case. For example, one of the attributes of sacrificial dilemmas is the number of harms prevented by acting. If the cognitive interference involves numbers, this may reduce attention to that particular dimension, unless it is so salient that it cannot be ignored. For example, Trémolière and Bonnefon (2014) used pairs of dilemmas differing in the number of people saved by killing one person. In general, time pressure or load reduced utilitarian responding when few were saved but seemed to increase utilitarian responding when the number saved was high. Subjects might pay more attention to numbers when they are large, i.e., give more weight to them when resolving a conflict between number saved and the repugnance of the action required to save them. Interference of any sort could primarily affect dimensions of the problem that are given less weight at the outset.

In sum, it seems that very little evidence supports any claim about the generality of effects across different scenarios, even scenarios of the same type (e.g., sacrificial dilemmas of the "Greene" type). And, if such general effects are found, they do not necessarily imply a sequential two-system theory. If some processes are more sensitive to cognitive resources than others, we might or might not want to say that they are part of different "systems", but that is an issue we need not address here. The methods we have used here might be applied to other areas where sequential two-systems theory seems to be supported. Perhaps moral judgment is an exception, and theory does quite well in other domains. (Hutcherson, Bushong & Rangel, 2015, make a similar argument to ours in the domain of altruistic choice.)

To summarize, our analysis yields three main conclusions. First, the prediction of the sequential two-system model concerning the RT difference when response probabilities are equal, holds for some studies but is reversed for others, yielding no overall effect. Gürçay and Baron (2017) explained such variation by suggesting that scenarios differ in the relative salience of information supporting one response or the other. The sequential aspect of the sequential model may apply to the order of attending to relevant pieces of information rather than the order of considering the responses themselves. Studies differ in many factors that could affect the former ordering, not only in the scenarios themselves but also in the way in which they are presented. Note that this sort of methodological problem cannot be overcome simply by the inclusion of additional scenarios and by statistical tests of generality across scenarios as well as subjects. (The latter tests are a minimal requirement, however, often ignored.) We conclude that the conflict model is generally a good description of how people respond to these dilemmas.

Our second conclusion is that utilitarian subjects take longer on the task. This is a small effect, which emerged by looking at all the studies together. We think it is related to the general finding that utilitarian subjects are more reflective. Baron et al. (2015) found that utilitarian responding, and utilitarian beliefs in general, were correlated with long response times, as well as higher accuracy, in solving problems in logic and arithmetic. We found a higher correlation with a measure of belief in actively open-minded thinking, a belief that, among other things, effective thinking requires searching actively for reasons why an initial or pet conclusion might be wrong, and treating such reasons fairly if they are found. Such search for reasons on the other side would make for longer response times, but response times alone could also be increased if people search for supporting reasons only. Thus, correlations with response time are fairly low and may be mediated by actively open-minded thinking.

Yet, our results (here and elsewhere) do not support the idea that additional reflection is specific to looking for reasons against an initial deontological conclusion. This idea would imply the sequential two-system model. Rather, the correlation between reflection and utilitarian responding may have more to do with the development of moral judgment in people before they come into the laboratory for experiments. This possibility is supported by correlations between utilitarian responding and measures of acceptance of utilitarian principles, or of beliefs about actively open-minded thinking (Baron et al., 2015, in press). It could be that certain kinds of education promote both utilitarian thinking and active open-mindedness. Or it could be that reflective thinkers, over time, gravitate toward utilitarian thinking as a result of questioning the deontological rules that they were taught. One result apparently at odds with the idea that individual differences are determined only by the prior history of the subjects, rather than the way they process information in an experiment, is the foreign-language effect of Costa et al. (2014), which suggests that differences in processing within an experiment may also matter.

A third, and weaker, conclusion results from our attempt to describe the data in terms of a drift-diffusion model. The results are consistent with the existence of "trial to trial" variability in drift rate, but we found no evidence that variability in the starting point is at all helpful in accounting for the data. Such variability would be present in our data (to some degree) if subjects differed in response bias toward or away from utilitarian responses. This negative result concerning starting-point variability needs to be followed up in other ways. It seems to suggest that individual differences in utilitarian responding are not due to individual differences in response bias but, rather, to differences in sensitivity to the different kinds of arguments for or against utilitarian conclusions.

References

Baron, J. (in press). Utilitarian vs. deontological reasoning: Method, results, and theory. In Bonnefon, J.-F., & Trémolière, B. (Eds.) *(Forthcoming). Moral inferences*. Hove, UK: Psychology Press.

Baron, J., Gürçay, B., & Luce, M. F. (in press). Correlations of trait and state emotions with utilitarian moral judgments. *Cognition and Emotion*.

Baron, J., Gürçay, B., Moore, A. B., & Starcke, K. (2012). Use of a rasch model to predict response times to utilitarian moral dilemmas. *Synthese*, *189*(Supplement 1), 107–117. (special issue on Psychological Models of (Ir)rationality and Decision Making, edited by C. Witteman & W. van der Hoek).

Baron, J., Scott, S., Fincher, K., & Metz, S. E. (2015). Why does the cognitive reflection test (sometimes) predict utilitarian moral judgment (and other things)? *Journal of Applied Research in Memory and Cognition*, *4*(3), 265–284. special issue on Modeling and Aiding Intuitions in Organizational Decision Making, edited by J. Marewski & U. Hoffrage. doi:10.1016/j.jarmac.2014.09.003

Bates, D. M., Maechler, M., Bolker, B., & Walker, S. (2015). lme4: Linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48.

Brown, S., Gretton, M., Heathcote, A., & Singmann, H. (2016). Rtdists: Response time distributions. r package version 0.6-6. http://CRAN.r-project.org/package=rtdists

Conway, P., & Gawronski, B. (2013). Deontological and utilitarian inclinations in moral decision making. a process dissociation approach. *Journal of Personality and Social Psychology*, *104*, 216–235.

Costa, A., Foucart, A., Hayakawa, S., Aparici, M., Apesteguia, J., Heafner, J., & Keysar, B. (2014). Your morals depend on language. *PLOS One*, *9*(4), 1–7.

Cummins, D. D., & Cummins, R. C. (2012). Emotion and deliberative reasoning in moral judgment. *Frontiers in Psychology*, *3*, 328.

Evans, J. St. B. T. (2007). On the resolution of conflict in dual- process theories of reasoning. *Thinking and Reasoning*, *13*, 321–329. doi:10.1037/xge0000107

Evans, J. St. B. T., & Stanovich, K. E. (2013). Dual-process theories of higher cognition: Advancing the debate. *Perspectives on Psychological Science*, *8*, 223–241, 263–271.

Greene, J. D. (2007). The secret joke of Kant's soul. In Sinnott-Armstrong, W. (Ed.) *Moral psychology. The neuroscience of morality: Emotion, disease, and development*, (Vol. 3 pp. 36–79). Cambridge, MA: MIT Press.

Greene, J. D., Sommerville, R. B., Nystrom, L. E., Darley, J. M., & Cohen, J. D. (2001). An fMRI investigation of emotional engagement in moral judgment. *Science*, *293*, 2105–2108.

Gürçay, B., & Baron, J. (2017). Challenges for the sequential two-system model of moral judgement. *Thinking & Reasoning*, *23*(1), 49–80. doi:10.1080/13546783.2016.1216011

Hutcherson, C. A., Bushong, B., & Rangel, A. (2015). A neurocomputational model of altruistic choice and its implications. *Neuron*, *87*, 451–462.

Kahneman, D. (2011). *Thinking fast and slow*. New York: Farrar, Straus, and Giroux.

Koop, G. J. (2013). An assessment of the temporal dynamics of moral decisions. *Judgment and Decision Making*, *8*, 527–539.

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review*, *85*(2), 59–108.

Ritov, I., & Baron, J. (1999). Protected values and omission bias. *Organizational Behavior and Human Decision Processes*, *79*, 79–94.

Rizopoulos, D. (2006). An R package for latent variable modelling and item response theory analyses. *Journal of Statistical Software*, *17*, 1–25.

Sloman, S. A. (1996). The empirical case for two systems of reasoning. *Psychological Bulletin*, *119*, 3–22.

Suter, R. S., & Hertwig, R. (2011). Time and moral judgment. *Cognition*, *119*, 454–458. doi:10.1016/j.cognition.2011.01.018

Trémolière, B., & Bonnefon, J.-F. (2014). Efficient kill-save ratios ease up the cognitive demands on counterintuitive moral utilitarianism. *Personality and Social Psychology Bulletin*, *40*, 923–930.

Trémolière, B., De Neys, W., & Bonnefon, J. F. (2012). Mortality salience and morality : Thinking about death makes people less utilitarian. *Cognition*, *124*, 379–384.

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software*, *36*(3), 1–48. doi:10.18637/jss.v036.i03